AALBORG UNIVERSITY
STUDENT REPORT

Aalborg University Copenhagen
A.C. Meyers Vænge 15
2450 København SV

Semester Coordinator: Henning Olesen

Secretary: Charlotte Høeg

Semester:
ICTE 4

Title:
Analyzing the potential of Differential Privacy in Data Sharing among Small and Medium-sized Enterprises.

Project Period:
Spring Semester 2022

Theme:
Innovative Com. Tech and Entrepreneurship

Supervisor(s):
Jannick Kirk Sørensen

Project Group: ICTE 4.2

Participant(s):
Iasonas Koniaris
Jens William Knudsen

Page Numbers: 51

Date of Completion:
September 27, 2022

Abstract

This paper examines the privacy protection concerns that discourages Small and Medium-sized Enterprises from sharing data. In order to tackle this problem we analyze the feasibility of using Differential Privacy as a means to alleviate these concerns. We find that Differential Privacy can be a useful and flexible tool, however the technology has not reached a suitable stage of maturity for commercial viability.

When uploading this document to Digital Exam each group member confirms that all have participated equally in the project work and that they collectively are responsible for the content of the project report. Furthermore each group member is liable for that there is no plagiarism in the report.

1 Introduction

The world of today views information almost as a commodity. While it does not have a physical form, information is processed, stored, and exchanged on a daily basis. Information is vital for enterprises and businesses in order to compete with one another [1]. Access to greater information often leads to a better product, better service, or even higher revenue. However, only a relatively small number of small and medium-sized businesses share data [2] [3]. While bigger enterprises have access to a huge pool of information, smaller enterprises have limited resources to work with. With initiatives such as the GDPR, companies must follow strict regulations when exchanging information. Furthermore, lack of trust and fear of exposing private information is halting enterprises from interacting with each other [1]. This leaves us with a market that can not be competitive in terms of information.

One of the technologies that have seen popularity in recent years is the concept of Differential Privacy [4]. Differential Privacy promises companies the ability to share or publish information while maintaining privacy. In this project, we will delve into the technicalities and possible applications of Differential Privacy with the aim of finding out if the technology can provide a safe and confidential way to exchange information. Our focus will be on small to medium-sized enterprises, and we will seek to discover whether Differential Privacy is the right solution to their data sharing problems

1.1 Problem Formulation

Our problem formulation consists of topics and themes that will be discussed and analyzed throughout the length of this report. The problem formulation that this project will be dealing with is as follows:

"Is differential privacy suitable for protecting privacy and confidentiality in data sharing among small to medium-sized enterprises?"

The report will also be dealing with a set of sub-questions that will be discussed before we can conclude on the problem formulation.

- How crucial is the need for small to medium-sized enterprises to protect their privacy?
- What are the problems of implementing differential privacy in data sharing?
- What benefits can differentially private data sharing provide?

1.2 Delimitations

This report will examine whether differential privacy can be considered as a solution for SMEs. The scope of the report however is to focus on a theoretical approach of how differential privacy can be considered as a solution. This means that implementations of code that depict the various algorithms and mathematical theories that differential privacy consists of will not be pursued in this report.

2 Methodology

This project began with the desire to help SMEs be more competitive relative to their larger counterparts. Our goal was to help SMEs in the area of data gathering. We quickly realized that data sharing among SMEs could be a useful tool to allow SMEs to gather larger quantities of data. This meant that our first step was to research the state of data sharing in SMEs and the problems that SMEs were facing when sharing data. We did this by first finding independent data sources on areas such as, how many SMEs were sharing data, what benefits they were receiving from data sharing, and the risk of data sharing. We also performed our own series of interviews to get a more qualitative perspective on the situation.

2.1 Process Model

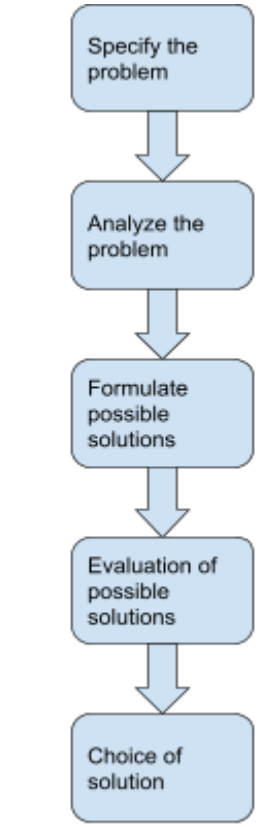
Our report will follow an analytical approach, therefore the framework and the overall process matches that of an analytical structure. The research throughout the report takes into consideration the carefully planned steps in order to ensure a qualitative outcome. After having explored the state of data sharing among SMEs and the barriers and benefits of data sharing, the results of which can be found in chapter 3, the next step was to identify a potential solution. This is where we discovered the concept of Differential Privacy. We identified Differential Privacy as a potential solution and began to explore the concept to gain an understanding of Differential Privacy. This is how we came to our problem formulation.

The next step was to analyze Differential privacy in regards to its benefits, problems, and how it may be used to help solve the problems facing data sharing among SMEs. As part of this analysis, we also investigated other methods of privacy protection which are currently being used by data sharing organizations such as anonymization and k-anonymity which can be found in chapter 4. We looked at the existing research into these methods to identify the strengths and weaknesses of the different methods, such that we could make an informed comparison.

The final stage was to bring our analysis together and discuss our findings. The final part of the project consisted of proposing how Differential Privacy may solve the problem but also its drawbacks, and concluding on our problem formulation.

This process followed a problem-solving strategy in order to tackle our problem formulation and the overall content of this project.

1. We need to specify the problem in order to be able to solve it. We use the Problem Formulation to define our problem and we use other sections such as Delimitations and Background to explain some of the concepts that we are going to examine. The problem formulation is created from brainstorming sessions that have been made in the early stages of the report. The idea behind this is to use the problem formulation as a guidance tool. This means that everything that is written throughout the life cycle of the project must have some purpose that serves the problem formulation. In our case small changes have been made during the development of this report to account for the changing circumstances.
2. We analyze the problem. We investigate what causes our problem to manifest and how different entities around our problem contribute to this manifestation.
3. We formulate possible outcomes and solutions to the problem at hand. We use the information and technology that we have gathered throughout our research to list potential solutions that would fit our problem.
4. An analysis is carried out on these possible solutions. We do this to determine how the solution tackles our problem. We look at the positives and negatives of each solution and how these affect the parameters that surround our problem.
5. For the last part, we choose and discuss a final solution for our problem and the project as a whole. The solution can have alterations and some recommendations even in this last part.



Five-step problem-solving strategy

2.2 State of the art

We use state of the art to review and analyze existing technologies and methods that have been used in the past and that are still used today. The methods in question are privacy enhancing techniques that can be used to mask private information when said information is published and in a manner that it can still be useful to the viewer. In addition to this, further technicalities are discussed while reviewing differential privacy as it is the main topic of the report.

2.3 Qualitative Research

Qualitative research is being carried out throughout the report. This type of research relies on data that can be acquired in the manner of documents, observations, questionnaires, interviews and other similar sources. While qualitative research is usually carried out in reports that revolve around sciences such as anthropology or sociology, here we approach the topic as a grounded theory. This means that we will collect data that are rich in character so that theories can be developed in both an inductive and deductive manner. In this report we collect papers, documents and conduct semi structured interviews in order to conduct our research[5].

2.4 Interviews

In order to accompany our Qualitative Research we constructed semi structured interviews to gain knowledge on the topic at hand. The nature of this interview is to allow the interview to divert into free discussion thus bringing new ideas to the table[6]. While semi structured interviews are used most of the time in social sciences, here we believe that it is a proper tool to examine the opinions and overall thought process of the interviewees. Specifically, one of the reasons that we use semi structured interviews has to do with the fact that we involve privacy as one of the main topics of this report. This is a sensitive topic that can generate various points of interest as interviewees might have different views on the matter.

2.5 Diffusion of Innovations Theory

One of the elements that we mention in this report is the Diffusion of Innovations Theory. The theory showcases how new practices, products and in general how new ideas adapt in society. There are several entities that take part in this theory, namely the innovators, the early adopters, early majority, the late majority and the laggards. In this report we use this theory to discuss the adoption rate of differential privacy if the technology should ever be implemented[7].

2.6 Network Effect

A Network Effect is a theory that is used in economics to explain the phenomenon that suggests that the value of a product or service increases when the number of people that partake in it increases as well[8]. The Internet is one of the most notable examples as in the beginning very few people were interested in it while nowadays it is part of our everyday lives. In our report we use this effect to describe the impact that differential privacy could make if implemented.

3 Background

The background chapter showcases the different aspects and entities that surround this project. Concepts such as SMEs, the EU, and data sharing will be explored and discussed. The chapter acts as a guide and as a way to inform and familiarize the reader with the elements that will be discussed and analyzed throughout the report. Furthermore, it analyzes some of these concepts to the extent of revealing how some of them affect the technology and the people around our problem.

In this project we are focusing on data sharing in SMEs. SME is an acronym for Small and Medium-sized Enterprises. We have chosen to focus on SMEs as the reports that we will be discussing show that they are not participating in data sharing as much, relative to larger enterprises. This is something we will be investigating throughout the background chapters. We will be investigating whether or not they really aren't participating in data sharing as much as larger enterprises, and what the barriers that might be holding them back are. But first we must define what we are talking about when we say SME.

3.1 SMEs

SMEs are essentially businesses where their revenue, number of employees, and overall assets are under a certain level. However, each country may have a different interpretation of what an SME is. Since we are exploring the European market the concept of SMEs will be interpreted as it currently is by the European Commission. The goal of the European Commission is to have a clear picture of what an SME is and thus avoid inconsistencies. According to the European Commission, an enterprise can be classified as an SME by the determination of two factors. The first is the staff headcount and the second how much the turnover is for every enterprise. Table 1 showcases the values in each column that describe the size of an enterprise. Note that in this project, we include the micro-enterprises under the SMEs umbrella.

Company category	Staff headcount	Turnover	or	Balance sheet total
Medium-sized	< 250	≤ € 50 m		≤ € 43 m
Small	< 50	≤ € 10 m		≤ € 10 m
Micro	< 10	≤ € 2 m		≤ € 2 m

SME Definition [9]

The European Union states that Small to Medium-sized Enterprises represent the backbone of the European economy as they consist of 99% of all businesses. It is estimated that 100 million people are employed by SMEs. Furthermore, the EU declares that SMEs present innovative solutions and combat challenges such as climate change. They even promote resource efficiency and provide social cohesion [10].

Now that we have a working definition of what SMEs are, we can start to look at whether or not SMEs participate in data sharing, what the drivers are and what the barriers are. We need to have an understanding of how SMEs interact with data sharing in order to understand whether or not differential privacy would be a good fit for them.

3.2 The EU and Data Sharing

According to the EU, SMEs are the backbone of the European economy that strives for information in order to maintain competitiveness [10]. In this goal however they are not alone as the EU is also interested in improving and assisting SMEs both economically and in terms of overall data. For example, as we mentioned in section 3.1, the EU has a set of parameters to determine if enterprises can be classified as SMEs. This allows national and communal measures to interact with each other when it comes to arranging and establishing measures, strategies, and even basic funds for SMEs.

This has been further approved by the European Investment Bank (EIB), the Member States, and the European Investment Fund (EIF). The notion of these bodies is that it improves effectiveness and consistency. It also limits the risk of distortion of competition. Once enterprises are eligible and fall under the SME category they can benefit from potential EU support programs. These programs can provide for example funding, fewer requirements, and even reduced fees [10].

Furthermore, there are European initiatives such as Data Pitch which depict the positives and advantages of data sharing to SMEs and other startups. This program managed to create 22,4 million euros worth of investments, sales and efficiencies, and it even created 112 jobs [3].

The EU also has an interest in data sharing between businesses. In 2017 the European Commission contracted Everis to conduct a study with the aim of deepening their understanding of data sharing and help contribute to the development of policy frameworks. This study was published in 2018 and can also help us gain insight into data sharing among businesses in the EU.

The EU report can provide us with information on data sharing in the EU, but there are a few things we need to keep in mind when evaluating its information in our context.
The report looked at data sharing in the EU in general. Our project focuses on data sharing in small to medium-sized enterprises, but the report also includes large enterprises. This means that the report can give us a general impression of data sharing in all sizes of enterprises. However, we need other sources to verify whether the EU reports information holds true for small to medium-sized enterprises or if the results have been skewed by the large enterprises. 32 of the 129 companies were large companies, while 35 were medium, 27 were small and 35 were micro.
The EU report also did not cover all types of enterprises. It focused on 6 sectors: “data-generating driving (i.e. automotive, transport and logistics), smart agriculture, smart manufacturing, telecom operators, smart living environments (i.e. home automation, sensors, robotics, or wearable technology), and smart grids & meters”[2]. This means that we can’t use this data to accurately conclude anything about enterprises outside these six sectors. Any conclusion on enterprises outside these six sectors would have to be extrapolation or supported by additional sources.

The EU report also specifically considered machine-generated data, which the report defines as “data produced without the direct intervention of a human by sensors or by computer processes, applications and services”. So if we are working with other information and data in addition to this type of data, then the results from the report may not be quite sufficient to cover all the types of data. However this potential broadening of our definition of data can sometimes skew the data in specific directions. For example, If some number of enterprises share machine-generated data, then the number of enterprises that share data, machine-generated or otherwise, could only be higher.

Another point about the EU report is that it did not distinguish between personal and non-personal data. This project focuses in large part on privacy protection and so distinguishing between personal and non-personal could be of interest, but since the report did not make this distinction we will need to find this information elsewhere.

3.3 SMEs and Data Sharing

The first step in understanding SMEs and data sharing is to look at whether or not they are sharing data.

The State of Data Sharing

The 2018 EU report on data Sharing found that nearly 4 in 10 companies are sharing their data and for 20% of these companies data sharing is their main economic activity[2]. At the same time 42% of companies in the study reuse data from other businesses. The EU report also cites the European Commission’s public consultation on Building a European Data Economy[11]. In this study, one-third of respondents shared some of their data, and more than half of the respondents were dependent on third-party data. This study suggests that data sharing is not rare, but still less than half of the businesses share their data.

Another study can help us gain further insight into how many SMEs share data. A British study published in April 2020 and conducted by the Open Data Institute and YouGov surveyed 2060 British businesses about data sharing[3]. In this survey, 8% of micro-businesses, 24% of small businesses, and 26% of medium-sized businesses answered “Yes. it does” to the question: “Does your business share any of its data with other organizations or individuals?”. This is in contrast to the large businesses where 43% answered yes.
The Survey also asked whether the businesses used shared data from other organizations. To this 16% of micro, 29% of small, 36% of medium, and 46% of large businesses answered yes.

A study performed by PricewaterhouseCoopers (PWC) in Germany and published in 2018 found that out of the 210 German companies 3 out of 4 companies exchange data[12]. But again the percentage of SMEs that exchanged data was smaller than among large enterprises. The percentage of large enterprises that exchanged data was 83% while the percentage of SMEs was 72%.

Growth
The EU Study on data sharing between companies in Europe also reports that “the percentage of data suppliers sharing data as their primary activity is expected to double in five years’ time”[2]. The study also reports that a third of the companies that are not currently sharing data see a possibility of sharing data within the next 5 years, and almost half of the companies that don't reuse data see a possibility of using data from other companies within 5 years. If these estimates hold true then they would indicate that data sharing will experience significant growth in the near future.

The Open Data Institute and YouGov study also suggests that there is at least some room for growth in data sharing[3]. One of the questions that were asked was whether or not increased data in their business sector would help their business grow. 33% of SMEs agreed or strongly agreed with that statement while 38% neither agreed nor disagreed. 24% percent disagreed or strongly disagreed and 6% didn't know. These answers are fairly spread out, but it does show that at least some of the participating businesses would be interested in access to more data.

The PWC study asked whether the businesses projected growth in demand for cross-company data exchange over the next 5 years. Out of the answers, 72% of SMEs projected an increase[12].

Data Sharing Partners
The EU study also reports that the main data sharers are large companies and that data sharing mainly occurs within the same sector. Businesses also prefer to share data with companies they have close business relations. This is supported by the Open Data Institute and YouGov study, where 44% of SMEs, who said they shared data, said that they shared data with business partners[3]. The second highest categories were customers at 37%, then regulators at 31%, then 22% answered national government bodies, and 22% answered industry bodies.
It is also interesting to see that 8% of respondents said that they shared data with competitors. This is a steep drop from the 44% who shared with business partners.
The places that SMEs receive data from are also similar as the top answers were business partners, customers, industry bodies, and national government bodies.

The percentage of SMEs that share data in Germany, according to the PWC study, seems relatively high[12]. However, when we look at who they are exchanging data with it becomes more clear. According to the study, 83% of the businesses exchange data with business customers, 53% exchange data with suppliers, while only 15% exchange data with competitors. This shows that exchanging data along the supply chain is common, but horizontal data sharing is far less common.

Shared Data
According to the EU study, the main type of data being shared is information from internal IT systems such as information about products, services, sales, logistics, customers, partners, or suppliers. IoT data is also commonly shared. Of this type of data real-time or near-real-time data is most frequently shared.
The study also suggests that companies only share a small part of their data.

Conclusions on the State of Data Sharing in SMEs
Now that we have gathered some information on the state of data sharing in SMEs we can begin to talk about how this might affect technologies such as differential privacy.
The first question we were interested in was whether or not SMEs were sharing data. From our sources we find that it varies depending on the source and the category of businesses you ask. This means we can't give a definitive answer as to how many SMEs share data. However, the data that we have gathered suggests that the SMEs that share data are in the minority. But data sharing SMEs are not an insignificant percentage of the total. Our sources also suggest that SMEs have an interest in sharing data and that the proportion of SMEs that share data is growing. This means that not only will differential privacy or any similar technology need to struggle with the adoption of that particular technology but also with the adoption of data sharing in general, in the large proportion of SMEs that don't already share data.

Data sharing within SMEs does seem to mainly be focused on data sharing between non-competing businesses. There is a portion of SMEs that share data with competitors but it appears to be significantly smaller. This means that a data sharing technology first and foremost needs to be able to share data between supply chain partners and other non-competing businesses and organizations. The technology should also be able to support data sharing between competing businesses but it appears to be a lower priority for SMEs.

3.4 Benefits of Data Sharing

To evaluate data sharing technologies we need to understand the benefits that SMEs want out of the technologies. We need to understand what SMEs want from data sharing, so that we know what to look for in data sharing technologies. If a technology is unable to provide the benefits that the SMEs desire or they need to be sacrificed in order to use the technology then the technology will not be suitable for this purpose. Any additional benefits that a technology may provide, that is not one of the desired benefits, may be useful but will not be as valuable.

The sources we have found suggest that one of the main benefits of data sharing is the strengthening of business relationships [3],[2], [12]. By sharing data and information with competitors, supply chain partners or other organizations the SMEs can build a better relationship with these other businesses. Strong business relations can have many benefits that can be helpful to the SMEs. Any data sharing technology needs to allow the business to build strong relations with other businesses.

Our sources also seem to indicate that another benefit of data sharing is the development of better products, services and business practices [3],[2],[12]. All of these state that businesses use data sharing to help them develop better services, products and business practices.

3.5 The Benefits of a Transparent Market

We have established that there is a desire to share information between SMEs and that they see a number of benefits in sharing data. Another benefit is that in the event that a data-sharing system is implemented for a number of SMEs then the market would in theory be more transparent. For example, if a small city adopts this system and if this city, for the sake of the argument, is not interfered with by other markets then it would in theory be closer to perfect competition. Perfect competition is a theoretical market structure that is characterized by attributes such as enterprises that sell an identical product, all enterprises can not influence the market price of the product, enterprises can enter and exit the market without cost, and more [13].

However, the system can only introduce one of these attributes of perfect competition and that is that enterprises have an abundance of or complete information over the market and its products. This statement is not by any means bulletproof, as even if the enterprises can have complete information over the market it remains up to the enterprises to harness or to overall interact with this information. Even if they do, Information is not completely transparent as enterprises might have some information that they do not want to share thus making this information not fully transparent. Lastly, we want to illustrate that although it is not possible to achieve true competition with the system, it is possible to achieve opposition to a market that has a monopoly. Overall it is fundamental that the system not only provides enterprises with information but that it actually strives for a more perfect competition between the enterprises [13].

3.6 Barriers to Data Sharing

We gather that data sharing and overall access to information is vital for SMEs. We also state that the EU is interested in assisting SMEs to achieve greater information-sharing capabilities. However, we must ponder why data sharing among SMEs has not seen higher implementation within the European communities.
The EU is currently in the motion of deploying regulatory measures that help share business-to-business (B2B) data. However, the notion is that companies are hesitant to these B2B solutions as they do not see value when asked to share their prized data [14]. Data can be used by many different entities and organizations at the same time. Meaning that several organizations can benefit from this without limiting each other. Usually, negative limitations are found in markets where enterprises bestow copyright claims and infringements [15]. As it currently stands enterprises opt to keep data to themselves instead of increasing their pool. One of the main reasons that enterprises are not sharing data is that they are genuinely concerned about privacy. Furthermore, there are other reasons such as lack of demand, lack of skill within the company, fear of giving out trade secrets, fear of misappropriation, and many others [2].

The legal aspects that accompany enterprises when they commit to data sharing practices are also one of the reasons why enterprises hesitate to exchange information [3],[12]. To start things off, data protection under GDPR states that there is a distinction between the different categories of personal data.

Enterprises may be able to process data such as names, addresses, income, and passport numbers but they are not allowed to process information such as ethnic origin, sexual orientation political opinions, or religious beliefs. An action that enterprises must take is that they have to appoint a data controller, this role is responsible for how personal data is processed.
An additional role that they have to establish is the data processor which is responsible for the storage and process of information. Another role that has to be filled by an enterprise is the role of the Data Protection Officer. The officer cooperates with the Data Protection Authority in order to guide employees on how to process and store information. Furthermore, the parties that partake in data sharing must sign written contracts that allow each other to process the exchanged information. Consent, contractual obligations, and legal obligations must be met by both parties before processing any information[16].However, we know that the EU is also trying to implement such regulations that increase data sharing and at the same time, it does so in an incentivizing way. This is where our project comes in to assist the EU and this new direction that is taking place.

3.7 Risk of Data Sharing

Now that we know that concerns about privacy and confidentiality are the barrier that prevents SMEs from data sharing, we will look at these concerns. We need to understand the potential risks of data sharing breaches so that we can evaluate any solution to this problem.
In a traditional risk assessment, we evaluate risk based on the severity of the consequences of an event and the probability of the event. However, a privacy protection system is supposed to protect data by enabling the creators of the system to better control the probability of negative events. This means we need to be looking at the potential consequences of the data privacy breaches so that we can better understand the appropriate level of protection and thus the acceptable probability of privacy breaches.
The first assertion we might make is that the probability of privacy breaches should be zero. The problem with this is that to achieve this probability, the system would have to be designed in such a way that it becomes non-functional, or loses all value as a data-sharing system.
This relation between privacy and utility will be discussed further in section 6.1 data privacy vs utility. We need to find the correct balance between functionality and risk, so we must discuss the consequences of potential privacy breaches.
Since this project is centered around data sharing between SMEs we can broadly categorize data into two types of data that need to be protected, private data about the customers and confidential data about the enterprises themselves. There can be different consequences for privacy breaches of each of these types of data.

First, we will look at the potential harm that can result from data breaches on private data about customers or employees. We can categorize the harm into 5 different categories: physical harms, economic or financial harms, mental or psychological harms, harm to dignity or reputation, and societal or architectural harms [17].

- **Physical Harms**

Physical harms include physical injury and even death. In one case the breach of privacy concerning a woman's workplace lead to her murder[Remsburg vs. Docusearch, Inc]. This is an extreme case, but it demonstrates potential harm that can result from a breach of privacy.

- **Economic or Financial Harms**

The subject of a privacy breach may suffer economic or financial harm from things such as fraud or identity theft.

- **Mental or Psychological Harms**

Subjects may suffer from mental or psychological harm due to a privacy breach. This can, among other things, arise from things such as fear of others using the data.

- **Harm to Dignity or Reputation**

Embarrassment or humiliation is categorized as harm to dignity or reputation and can be caused by a subject's private data being exposed.

- **Societal or Architectural Harms**

The final category includes the effects of monitoring, social control, and the effects on free speech and civic life.

Confidential information concerning the enterprise itself can also be compromised and lead to different types of losses.

- **Loss of Trust**

If an enterprise's confidential data is breached it may cause its business partners to lose trust in the enterprise.

- **Loss of Reputation**

If an enterprise has its confidential information breached, it may suffer a loss of reputation. Customers may think twice about doing business with the enterprise.

- **Loss of Trade Secrets**

With the loss of trade secrets, an enterprise can lose its competitive advantage in the market as the trade secrets can become public.

- **Financial Loss**

The aforementioned losses that can occur due to a breach of confidential data, and can carry with them financial losses. A loss of trust, reputation or trade secrets can easily be accompanied by financial loss as the enterprise loses business partners, customers, and market advantage.

Harms or losses to the subjects of the data that has been compromised are not the only consequences of a data breach. There can also be legal consequences. As an example, the EU has instituted fines through the General Data Protection Regulation(GDPR). These fines can be as high as 20 million euros or 4% of total global turnover, whichever is highest[17], [18].

As we can see there's a large variety of different consequences that must be considered whenever we are dealing with data protection. However, there's also a large disparity between the severity of the consequences. The potential negative consequences will largely depend on the specific data that is compromised, and so different levels of protection must be implemented when there are different potential consequences of protection failure.

4 State of the art

We have established what SMEs are and the state of data sharing among SMEs. We have seen how they benefit from data sharing as well as the problems they are facing. We have found that the main barrier to data sharing appears to be privacy concerns. If data sharing among SMEs is to grow further this problem of privacy must be solved. Therefore we will now discuss the potential solutions that may be able to solve these problems.

We will first discuss the well-known and simple method of anonymization, then we will explore the more advanced anonymization technique k-anonymization, and finally explain the newer and less widespread method of Differential Privacy and how it stands out. Our main focus will be Differential Privacy as it is a newer and lesser known technique.

4.1 Anonymization

One way systems have tried to protect businesses and people from the harms of data breaches is anonymization. By simply removing any personal identifiers from the data, it is thought that the data becomes anonymous. Without these personal identifiers, the data can be used and shared without risking harm to the data subjects. Unfortunately, this is not true [19]. Even if data such as names or IDs are removed from a data set it may still be possible to identify people in the data through methods such as de-anonymization [20].

As an example in 2006, Netflix released a data set of 100 million user ratings, as part of a competition to improve their recommender system [[21]. These 100 million user ratings had been anonymized. Netflix believed that it would prevent people from identifying the users. However, a research team was able to compare the Netflix data set with a data set from the Internet Movie Database (IMDB) [22] and identify large numbers of users [20].

These de-anonymization, or linkage attack, methods take a data set and compare the entries in the data set to the entries in some external data set. They then look for entries that match in both data sets. The more two entries match up the higher the probability is that they belong to the same data subject. This method can’t guarantee that the two matching entries are based on the same data subject, but the probability can become so high that the subject is effectively identified.

With these types of linkage attacks, it is clear that simple anonymization isn’t enough to protect privacy. We need to find privacy protection methods that go beyond simply anonymizing data.

4.2 k-anonymity

The concept of k-anonymity is used to disclose person-oriented data structures with the protection of privacy in mind to whoever is interested in it. It is done without making it possible to identify the individuals that are within these data sets while still maintaining the usefulness of the data. The definition states that “Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals”. [23] In order for the released data to have the k-anonymity model applied, the information connected with each person must not be distinguished by k - 1 peoples data that appear in the same overall data release. The more comprehensive statement is written as such, “*Let T (A1, . . . , An) be a table and QI be a quasi-identifier associated with it. T is said to satisfy k-anonymity wrt QI if each sequence of values in T [QI] appears at least with k occurrences in T [QI]*.” [23] There are a few ways to apply k-anonymity to a given data set. The first method for applying k-anonymity is the method of generalization. **Generalization** is a method where values that represent individuals are exchanged for broader categories of values. For example, if you have the year of birth of an individual that was born in ‘1995’ you can replace it with the broader attribute of being born between ‘1990’ and ‘2000’. **Suppression** is another method for achieving k-anonymization. In this case, you suppress information by removing attributes as a whole. To give an example we take the attribute of an SME which states what kind of an SME it is, for example, a cafe, a restaurant, or a cinema and we replace it with a null value. A full example is shown in table [link] where attributes of SMEs are shown before and after the implementation of k-anonymity by inserting Suppression and Generalization methods into the data set. The names of the SMEs and their owners' names were suppressed, while the year of establishment and the number of staff were generalized.

Table 4.2.1

Name of SME	**Year of establishment**	**Name of owner**	**Type of SME**	**Region**	**Number of staff**
CafaLusso	2010	Carla	Cafe	Copenhagen	4
Jibbers	2018	Bob	IT company	Zealand	7
Qaffo	2020	George	Restaurant	Northern Denmark	12
Voile	2020	Michael	Bar	Zealand	8
Kairu	1995	Jane	Hotel	Southern Denmark	10
The Hive	2000	Katherine	Computer Hardware company	Copenhagen	3
Kaleido	2006	John	Jewelry store	Central Denmark	2

Table 4.2.2

Name of SME	Year of establishment	Name of owner	Type of SME	Region	Number of staff
Not Disclosed	2005-2015	Not Disclosed	Cafe	Copenhagen	1-10
Not Disclosed	2010-2020	Not Disclosed	IT company	Zealand	1-10
Not Disclosed	2012-2022	Not Disclosed	Restaurant	Northern Denmark	10-20
Not Disclosed	2012-2022	Not Disclosed	Bar	Zealand	1-10
Not Disclosed	1990-2000	Not Disclosed	Hotel	Southern Denmark	10-20
Not Disclosed	1990-2000	Not Disclosed	Computer Hardware company	Copenhagen	1-10
Not Disclosed	2000-2010	Not Disclosed	Bar	Central Denmark	1-10

While the tables above are used as an example of how k-anonymity can be implemented there are some problems with the concept when someone wants to commit “attacks” against k-anonymity.

Unsorted Matching Attack: This type of attack refers to the tuples that the data set or table depicts[24]. The order of the tuples can be assumed by the attacker and in a real-world scenario it often is a problem as sensitive data can be leaked or altered. However, it can be solved by randomizing the order of the tuples[25].

Complementary Release Attack: Typically a common theme with these data tables is that when they are released, the attributes of the table constitute the quasi-identifier thus making them a subset of the attributes within the table. This means that when the data table is shown with the k-anonymity properties implemented it should be viewed as a combination of other external information. Furthermore, when there are more releases of the same data, it must take into consideration all of the released attributes of the quasi-identifiers in order to prevent linkage attacks with the original data table[25].

Temporal Attack: Sometimes that data tables or sets that are released go through dynamic changes. This means that if a given table that is released goes through some changes, for example, if some tuples are added or changed altogether, the overall data that has been released so far can be temporal inference attacked. If we release table T1 and add later in time some new tuples into the table making it effectively table T2, there are no guarantees that the previous version of table T1 is linked to the new version of table T2. This leads to sensitive information being leaked[25].

k-anonymity Mitigations: There are several drawbacks to k-anonymity however there are a couple of methods that can enhance the effectiveness of k-anonymity. The l-diversity method is an additional step imposed on k-anonymity in order to ensure a better quality of anonymization. l-diversity acts as a benchmark, it is used to ensure that anonymization has been implemented to a level where re-identification can be avoided. There are many algorithms that can achieve anonymization using this method as a benchmark. However, it cannot account for background knowledge attacks. t-closeness can be used as a further enhancement of l-diversity to avoid such attacks. Nevertheless, the more of these methods that are applied to a data set the less effective the information will be for the user[26].

4.3 Differential Privacy

When dealing with privacy in data sets there's a variety of methods that can be used, many of them use the concept of differential privacy. Differential privacy is specifically concerned with how to keep the data set useful in analysis while preserving the privacy of individuals in a data set. However, overly accurate answers to too many questions will still ruin the individual's privacy [27].

The core promise of differential privacy is that any individual data subject will not be affected by allowing the data to be used for study or analysis. Any individual in the data set should be able to participate or not participate in the data set, and not be affected either way, while the study or analysis of the data set should still give the same result.
An example that is often used to explain differential privacy is the example of a smoker participating in a study. The study may show that smoking causes cancer, thus raising the smoker's insurance premiums. In this example, differential privacy takes the stance that the smoker's privacy was not compromised because the impact on the smoker would have been the same whether or not they participated in the study. It is the result of the study that affected the smoker, not their participation [27].
Differential privacy is defined by the difference between the two outputs of an algorithm given two data sets, with or without one individual. This difference in output is captured by the privacy parameter ε, the smaller the ε the better the privacy. However, differential privacy is a definition of privacy, not an algorithm. There are many different algorithms for different tasks and values of ε.

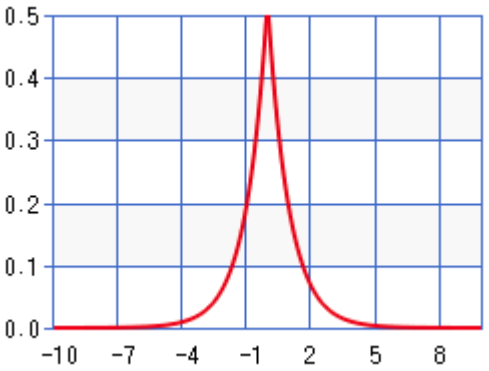
4.4 Adding Noise

Differential privacy works by adding noise to keep individuals' data private. By taking the real answer to a query that is made on the data set and adding noise, it can obscure the true values in the data, and thus protect participants' privacy. But this leaves the question of how much noise should be added? If too much noise is added then the accuracy of the query result will suffer, but if too little noise is added then the privacy will suffer.
A good way to add noise is to draw from the Laplace distribution [27]. The gaussian distribution can also be used, however the Laplace distribution has the advantage of being narrower while still having the probability of large variations. The method of extracting the amount of noise from the Laplace distribution is called the Laplace mechanism.

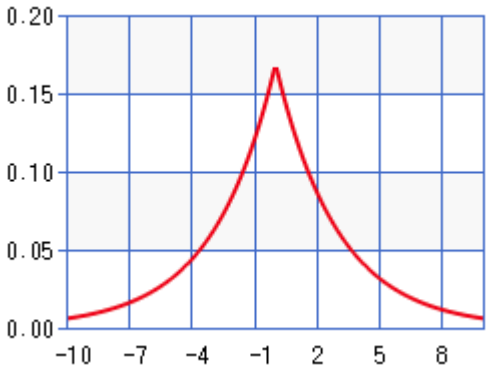
The Laplace mechanism calculates the noise that will be added by drawing a random value from the Laplace distribution. The Laplace distribution is defined by a center point and a scale factor. The center is placed at 0 so that it provides both positive and negative values. The scale factor is used to determine the magnitude of the value, a higher scale factor means a higher likelihood for high values of noise. The scale factor for the noise will be calculated based on the sensitivity divided by the ε value, Δf/ε. In the following two sections, we will discuss these two values and how they can be chosen.

4.5 Function Sensitivity

Differential privacy provides privacy by giving plausible deniability to individuals as to whether or not they are in the data set or whether or not their data is correct. Anyone querying the data should not be able to determine whether any one individual has been added, removed, or has changed their data. However, the addition, removal, or change of any individual data point, has different effects on different query functions. The sensitivity of a function is used as a way of describing how much the result of the function changes based on the input. The function sensitivity is used to determine the magnitude of the noise that needs to be added. The sensitivity of the function is used to define the Laplace distribution that noise will be drawn from. The Laplace distribution is a symmetric version of the exponential distribution and is defined by a center and a scale factor. By placing the center at 0 it can create both positive and negative values as the noise, and the scale factor determines the width of the distribution.



Laplace distribution with scale factor 1



Laplace distribution with scale factor 3

As the scale factor increases the probability distribution widens, resulting in a higher probability of getting larger values. The sensitivity of the query function is used to determine the scale factor in combination with ε.

A simple example of how different functions have different sensitivity is the two following equations.

$f(x) = x$

This equation has a sensitivity of 1 since changing x by 1 also changes $f(x)$ by 1.

$f(x) = x * 5$

Whereas this equation has a sensitivity of 5 since changing x by 1 changes $f(x)$ by 5.

We can use this to start looking at any query function that maps the data set to real numbers.

The most simple of these functions is the counting query. Counting queries count up the number of entries in the data set that fulfill certain criteria. Counting queries always have a sensitivity of 1, as adding, removing, or changing a single entry can at most change the result by 1.
Another simple type of query is the summation query. The summation query sums up the values of some number of data entries. The summation query has no clear sensitivity as there is no set range of values that will always be summed up. The range of possible values depends on the type of data that is being queried. If we use people's ages as an example then the range could be 0 - 100, but there's no guarantee that someone older than 100 won't be added to the list. We can try to set a bound for the range of values, but there's no guarantee it won't be violated. As such summation queries are unbounded, and it is up to the individual implementation to set a reasonable bound. A reasonable bound is important, as a bound that is too narrow can be dangerous, as it would not provide the desired sensitivity for those that fall outside of it. But a bound that is too wide would cause the accuracy of the query to suffer. This problem of unbounded sensitivity is not unique to summation queries, there may be other queries whose sensitivity is also unbounded, these queries must also have a bound imposed on them.
Now that we have these two types of queries we can use them to build other queries. For example, a query asking for the average of a set of values is simply the result of the summation query divided by the result of the counting query.

4.6 Privacy Parameter ε

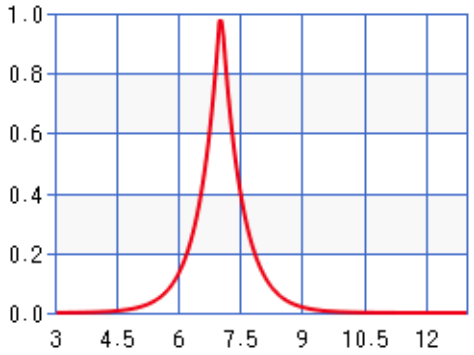
the second parameter used to determine the shape of the Laplace distribution is ε.
This parameter is used to set the desired probability that an attacker has of guessing the answer to a query without the noise, thus potentially breaching the user’s privacy.
Let's use an example of a database and query to demonstrate how ε can be determined. As an example, we take a database of four bars with their names and the number of different beers they serve.

Name	Number of beers
A	12
B	10
C	7
D	4

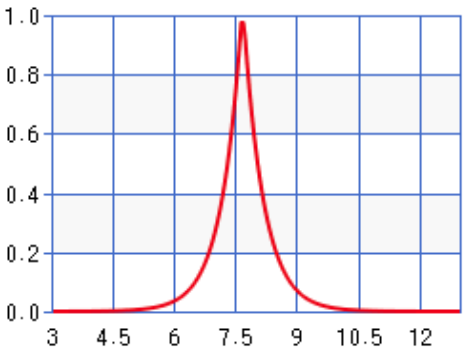
We want to release the average number of beers of some subset of bars, perhaps 3 of them have a special deal with a brewery. However, we don't want anyone to be able to tell which one of the bars is left out. Let's assume the worst-case scenario where the adversary knows all data, except which of the bars have special deals. The adversary knows both the names and number of beers of all four bars. Without performing the query the adversary has a 25% chance of guessing which of the bars does not have a special deal. However, after the query, if no noise is added, the adversary has a 100% chance of guessing correctly.

In this example, the adversary can simply calculate the average number of beers for each possible combination of bars, and find the combination that corresponds to the result of the query. But more generally the adversary can calculate the probability of all possible database states, and then once the database has been queried, update the probability of each state based on the query result. In our example, the updated probability will leave only one possible combination with 100% probability. This changes when the noise from the Laplace distribution is added.

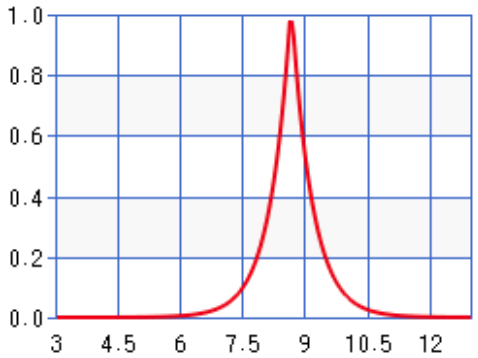
The adversary can create a series of Laplace distributions, one for each possible combination of bars. These would be the following distributions:



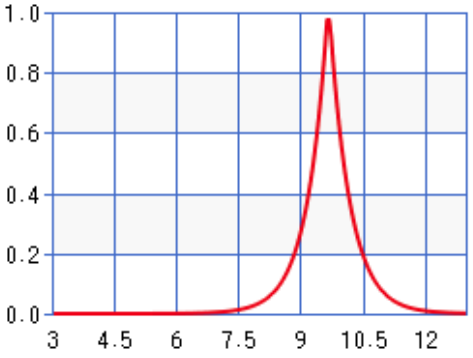
Laplace distribution location: 7,00 scale:0,5



Laplace distribution location: 7,66 scale:0,5



Laplace distribution location: 8,66 scale:0,5



Laplace distribution location: 9.66 scale:0,5

Each Distribution is centered on the average number of beers of the corresponding combination of bars. The scale factor is the scale factor employed by the differentially private mechanism.
As discussed previously the scale factor is the function sensitivity divided by ε. In this example, the scale factor is set to 0,5. The adversary can then perform the query and compare the likelihood of the query result coming from each distribution. If the query result is 9, the probability of this value originating from each distribution can be calculated using the Laplace probability density function:

$$f(x, a, b) = \frac{1}{2*b} e^{-\frac{|x-a|}{b}}$$

x, is the query result.
a, is the location parameter.
b, is the scale factor.

Bar combination	Average number of beers	Probability of query result: 9 given this combination
B,C,D	7	0,0183
A,C,D	7,66	0,0685
A,B,D	8,66	0,5066
A,B,C	9,66	0,2671

From this we can see that the most likely result is the bar combination; A,B,D.
We can also calculate the confidence in each combination.

Bar combination	Confidence
B,C,D	0,02
A,C,D	0,07
A,B,D	0,58
A,B,C	0,31

This means that given the query result of 9, the adversary can be 58% confident in guessing the bar combination A,B,D.
However, this confidence falls when the scale factor of the Laplace distribution increases. As the distributions widen, the overlap between them increases. With a larger overlap, the likelihood of the query result originating from either one of them converges. As the scale factor approaches infinity the distributions become infinitely wide and thus indistinguishable. If the distributions become indistinguishable then the probabilities are the same for each distribution. This means that confidence in any combination is the same. In our example, we have 4 possibilities which mean, the confidence in any possibility would be 0,25. If there were 5 possibilities, then each would have 0,20. This sets our lower bound as one divided by the number of entries. This means the more data in the database, and thus possible query results, the lower, the lower bound is.
Differential privacy uses the function sensitivity divided by ε as the scale factor. As ε approaches 0 the scale factor increases and the adversary's confidence approaches the lower bound.
This method also allows us to work backward and find the proper value for ε given a desired maximum adversary confidence. If we decide on a maximum value for the adversary’s confidence, we can use the data set being queried and the query functions sensitivity to find the corresponding value for ε.

5 Problem Analysis

In this section we will be analyzing the problem further to get a more detailed and in depth understanding of the problem. We will be doing this by performing our own interviews and by using the diffusion of Innovations theory.

5.1 Interviews

To get a deeper and more qualitative understanding of data sharing in SMEs we conducted a series of interviews with SMEs.
The goal of the interviews was to shed light on data sharing in SMEs. We focused on how businesses were sharing data, what kind of data, and how it was shared. We also asked what benefits they gained from sharing data and if they could see the potential for further benefits from sharing more. The barriers stopping SMEs from sharing data or sharing more data were also of interest so that we can understand the problems SMEs face when sharing data.

These interviews were semi-structured. The interviewer would ask about certain topics to ensure we gained the information we were looking for. But the interviewer would also dive into the interviewees' answers, ask clarifying questions and probe further.
The interviews were conducted in two separate areas, Copenhagen Denmark, and the Greek island of Samos. 7 interviews were performed in Greece and 6 were performed in Denmark. These two different locations allow us to get a broader look at SMEs due to the fact that they are so different. Copenhagen is a large city in northern Europe where there is a large number and variety of businesses. The small Greek island is a much smaller community with less competition. These factors may change the SMEs' stance on data sharing.

The interviews were performed in either Greek, Danish, or English to accommodate the interviewee. This was done to try and make the interview more approachable and easier for the interviewees to participate and communicate. This does however mean that there may be differences in the interviews due to translation. However these interviews were trying to explore and understand the interviewees, so hopefully, the deeper explanations were able to compensate for translation differences.

The businesses that were interviewed in Greece were a variety of different businesses including a book store, a cafe, a cafeteria, an IT and software consultancy, a jewelry shop, a cinema, and a tourist shop. The SMEs interviewed in Copenhagen were all some variety of restaurant businesses. The lack of diversity in the Copenhagen sample is in large part due to the large proportion of SMEs in Copenhagen being restaurants. Many other types of businesses in Copenhagen were too large or were unavailable for us to interview.

List of Interviews

Interview	Location	Date	Duration
1	Samos	July 7 2022	07:12
2	Samos	July 8 2022	24:05
3	Samos	July 14 2022	07:52
4	Samos	July 14 2022	03:22
5	Samos	July 15 2022	03:23
6	Samos	July 23 2022	06:04
7	Samos	July 23 2022	05:23
8	Copenhagen	July 25 2022	07:45
9	Copenhagen	July 27 2022	05:23
10	Copenhagen	July 26 2022	04:15
11	Copenhagen	July 27 2022	08:18
12	Copenhagen	July 27 2022	03:25
13	Copenhagen	July 29 2022	09:44

The State of Data Sharing

Part of our interviews with SMEs was focused on answering the question of whether or not they shared data. The results we got were that 5 out of 7 of the Greek SMEs shared data while 4 out of 6 Danish SMEs shared data. However, the type of data sharing varied a lot. Several of the SMEs shared data exclusively for the purpose of receiving a service or product in return. For example, some SMEs used an external booking service. These booking services were provided by another business and required them to share data in order for the service to function. The proportion of interviewed SMEs that shared data that were not directly related to the use of a service or product was significantly lower. Only 3 out of 7 Greek SMEs and 1 out of 6 Danish SMEs were sharing data for the purposes of sharing and gathering data. These SMEs would share data and information in order to receive information that they themselves could then act upon. For example, the cinema would share data on the success of their movies with other cinemas to get a better idea of the overall success of the movies. This data could then be used to strategize and improve their future success.

While only a small fraction of the interviewed SMEs shared data, some of the SMEs that did not partake in data sharing were still interested in receiving data and information from other businesses. One of the SMEs even referred to it as “spying”. They described going to their competitors and looking at their store, their products, and their prices.
This information is of course all publicly displayed. This shows that many of the SMEs are interested in receiving data and information but have concerns about sharing their own.

Data Sharing Partners
From our interview, we also got an insight into who SMEs are sharing data with. One of the categories of other businesses that SMEs are sharing data with is supply chain partners. Sharing data with supply chain partners allows the SMEs to express their product needs. But the main data sharing partners were other businesses that required the exchange of data in order to provide a service to the SME.

Only 4 out of the 13 interviewed SMEs shared data with competitors or similar businesses. Some of the cited reasons for not sharing data with competitors were privacy concerns, suspiciousness, and previous bad experiences. The results of our interviews would seem to agree with the EU report, that SMEs prefer to share data with businesses that they have close business relations with.

Shared Data
One of the types of information that our interviewees shared was information and ideas about products and services. This type of information can include some quantitative data but seems in large part to be qualitative data. The SMEs talked about sharing and being interested in how competitors were running their businesses and ideas for their products and services. This type of information may be difficult to share in a rigidly structured format. Some of the SMEs stated that they would share this type of data verbally, or would go look at how competitors were doing.

However, there was also some quantitative data being shared. Data such as the prices of products bought from their suppliers and the prices of the products sold to their customers. Another type of data being shared was employee wages. These types of quantitative data could more easily be shared in a rigidly structured format, and statistical analysis tools would allow SMEs to more easily analyze this type of data.

Customer data was also one of the types of data that some of the interviewed SMEs either were sharing or would like to share. But when talking about this type of data the SMEs also stated their concerns about the privacy or security of customer data. It seems that the SMEs wanted to ensure the privacy of this type of data before sharing it.

Sharing Methods
Our interviews found three main ways that SMEs shared data, Verbal, organized sharing, and automated systems.
Some of the Interviewed SMEs described talking with other businesses and sharing information this way. This was described as a more causal method of sharing data.
Another way that some of the SMEs shared data was through industry organizations or events. One SME mentioned an expo where businesses would share information.
The final method was through an automated system. This method however, was mainly used to share information with businesses that provided some type of service like a booking system to the SME.

Conclusions on the Interviews
Much of what we found in our interviews supports the studies that have been done and which we have gone through in our background chapter. The proportion of SMEs that share data with other SME to gather data that they can then use is the minority. However we also found that many of the SMEs were interested in gathering more data, but many had privacy concerns or other obstacles preventing them from doing so.
The interviews also supported the EU reports claim that SMEs prefer to share data with close business partners, one of the reasons for this is that data sharing helps SMEs build stronger business relations.

Another benefit we have found both in our background research and our interviews is that data sharing can help SMEs develop better products and services. When talking to the SMEs some of them mentioned how they used data, information and ideas from other businesses to improve their business. This category can be difficult to quantify, as to what exactly is needed of the data sharing technology to provide this benefit. However, some of the things mentioned by our interviewees was the sharing of prices and wages, so that they could be more competitive and be able benchmark themselves relative to other businesses. This means that a data sharing technology needs to allow for the sharing of this type of data. The sharing of ideas was also mentioned, which may be a more qualitative type of data. This means that a data sharing technology should ideally be able to share both quantitative and qualitative data.

Our interviews also shed more light on the barriers to data sharing. When we performed our own interviews with SMEs we found evidence to support the studies which we have explored earlier in the background chapter, and their claim that legal and privacy issues are a big barrier to data sharing. 10 out of the 13 SMEs we interviewed named privacy, security, and legal issues as some of the barriers to data sharing. This shows that these issues are one of the main barriers to data sharing and must be dealt with if data sharing is to grow.

Another barrier we identified from our interviews is a lack of resources to dedicate to data sharing. Two of our interviewees said that they did not have the necessary resources to dedicate to data sharing and another mentioned a lack of digitization. The small number of SMEs that mention this as a barrier fits with some of the other sources we found. In the Open Data Institute and YouGov survey only 8% of SMEs said that “Lack of technical in-house skills/ capacity” was a barrier to data sharing. This is a very small percentage of the SMEs. The technical and resource barriers to data sharing do not appear to be a big barrier to data sharing, but it is a barrier for some. Data sharing technologies could benefit from being easy to use and not very resource demanding, but it should not be a high priority.

Additional Findings
An interesting observation we can make is that a larger proportion of the Greek businesses interviewed shared data compared to the Danish businesses. This may be because of cultural differences, the smaller community, or simply due to the smaller sample size. There are some interesting details that were revealed to us through some of the interviews. One of the people that were interviewed has a company that acts as an IT and software consultant for other enterprises. They mentioned that they are not seeing a full implementation of GDPR by a lot of companies.

In addition to this in another interview, a bookstore owner was saying that the recent introduction of digitalization in many state agencies has led to an increase in data sharing between them and their customers. However, it is problematic for the bookstore owner, the increase in data sharing has brought the burden of devoting time to helping customers with their everyday digital tasks. Until recently Greece has carried out all its state agency work through physical paper.

The sudden shift to the digital world has left many without the knowledge of how to do simple tasks such as accessing documents or sending emails. A typical task for the bookstore owner may be that someone comes in one day to print some digital files. The bookstore owner will have to devote time to help them locate their files and then send them to their bookstore email address in order to print the files. Out of the 10 minutes that the whole interaction will take the bookstore owner will not be paid for the service but for the number of papers printed which can amount to as little as a couple of cents.

An assumption that can be made is that the recent digitalization of the state agencies in Greece as whole makes it difficult for its citizens to adapt to the digital world. In addition to this another assumption could be that whatever data sharing technology would be used in the end it would have to be easy to access and to understand.

5.2 Adoption Rate and Network Effects

A subject that needs to be addressed is the willingness of enterprises to adapt to the system. If first and foremost they want such a system then a follow-up question would be as to when a full adaptation will commence by these enterprises and the markets that they are in. Diffusion of Innovations Theory is able to hypothesize how such an adaptation can take place. It showcases how technological, economical, and other advancements are integrated within societies. It starts by describing the categories of people that adopt an innovation. The Innovators are the first to adopt the innovation and they are willing to even risk their enterprise in order to try the new innovation. Next, we have the Early Adopters, this group is interested in new innovations and wants to try them out. The Early Majority are using said innovations and are a segment of the overall population. The Late Majority follows the early majority in integrating innovation as part of their lives. Lastly, we have the Laggards, these are the people that will be the last to integrate innovations into their lives or sometimes they will even choose not to adopt them at all [28].

The innovator’s and early adopter’s willingness to take risks would be an important part of the adoption of a differentially private data sharing system. Because of the relation between accuracy and protection in differential privacy, the innovators and early adopters would have to contend with either lower accuracy or privacy protection. Their willingness to take risks may allow the system to use lower levels of privacy protection early in the system's adoption and then increase the protection as more SMEs adopt the system. This would give a natural progression of lowering the risk as more risk-averse SMEs adopt the system. This positive feedback loop of lowering risk as more risk-averse SMEs join could help grow the system.

There are several factors as to why, how, and when a society adapts to new technology. These factors can be population, education, development, industrialization, and many more. In our case, one of the main struggles that the system can face is the time of the adaptation of this new system. We can not estimate how many enterprises will be willing to adopt the system. Furthermore, the system is inherently a network. As the system connects enterprises by sharing information with one another, the network will be limited by the number of enterprises that are willing to adopt it. The value in data sharing is the data being shared. This means that the value of a data-sharing system grows with the volume of data being shared. But this also means that if there are very few SMEs sharing their data the value of the system is correspondingly low. The first SME to adopt the system would gain no value from the system until other SMEs join the system.

Because of the initially low value of the system, there needs to be some incentive for SMEs to adopt the system. One incentive would be the future potential of the system. If SMEs see a large enough future potential for the system then they may be willing to adopt the system early. Another way to incentivize early adoption is through external incentives. As we have discussed the EU has shown an interest in business-to-business data sharing in Europe and may be willing to provide some additional incentives or encourage early adopters. Such support could be vital to the adoption of the system, especially because of the early system's combination of low accuracy, high risk, and low volume of data. Industry groups and communities can also incentivize SMEs to adopt the system.

6 Technical Analysis

Now that we have described how differential privacy works, we proceed to the analysis of whether or not differential privacy would be suitable for protecting privacy in data sharing among SMEs.

6.1 Data Privacy vs Utility

"Data Cannot be Fully Anonymized and Remain Useful." [27]
This means that any data relating to a subject cannot be used and at the same time remain private. The utility and privacy of data have an inverse relationship. We cannot have the same data be both perfectly private and useful, there must exist a tradeoff.
As an example, if a business wishes to keep its desire to purchase a product entirely privately, then it cannot purchase that product. To purchase a product they must at some point reveal their intent to purchase the product to a seller. The more sellers they reveal their desire to purchase the product to, the more options and the more competition for the sale there will be.

Differential privacy is a definition of privacy that attempts to work with this tradeoff. An algorithm that analyzes a data set is differentially private if you cannot tell whether an individual's data was included in the data set or not, based on the result [29]. We can capture the difference between the results with or without any particular individual with the parameter, ε. A smaller ε provides more privacy but lower accuracy of the result and thus lower utility [27]. Differential privacy tools are designed to provide a certain value of ε, thus providing a controlled trade-off between privacy and utility.

As we have shown, SMEs can benefit from data sharing but they are concerned about the privacy risk. Differential privacy does not allow us to simply remove the risk but allows us to make a controlled trade-off between privacy and utility. Differential privacy is not a perfect solution, if it is implemented it will diminish the benefits of data sharing. This leaves the open question of whether or not SMEs would still be interested in data sharing, if the benefits are reduced in exchange for lower risk.

This question may be answered differently from enterprise to enterprise, depending on how big the benefits and risks are to them individually. Some SMEs may be able to better utilize the shared data based on analytical skills or how applicable the data is to the specific SME. Likewise, the risks may vary from SME to SME. SMEs that deal with more sensitive data may feel the risk is higher than those that deal with very little sensitive data. As we have mentioned, SMEs make up a large percentage of EU businesses and so, we will potentially be dealing with a large variety of differing opinions on whether or not the loss of utility in exchange for privacy protection, still makes data sharing favorable.
Something noteworthy is that a system such as this that enables SMEs to exchange information can be implemented according to the needs of the SMEs. To give an example the system does not necessarily have to cover or be implemented for the whole European Union. It can be split into smaller sections. These sections can be Countries, Communities, and even economic sectors. These groups can be responsible for their own rules, laws, regulations, and overall implementation of this data-sharing system.

We have also described how data sharing can help create a more transparent market, and move it closer to a situation of perfect competition. However, differential privacy does not make the data completely transparent. Differential privacy introduces noise to any query made on the data. This means that each SME would not have access to a complete and clear overview of the market. However, data sharing with differential privacy would help to provide a more transparent market, just not a completely transparent market.

6.2 Data Storage and the Injection of Noise

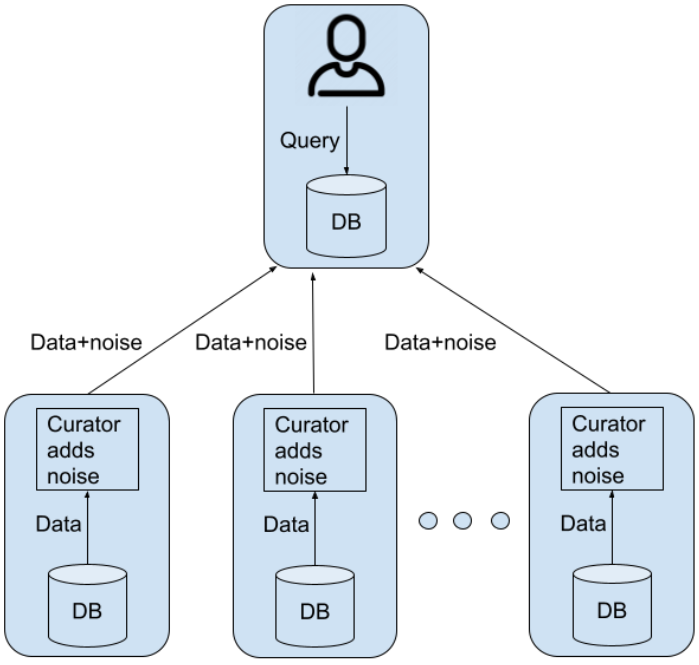
When applying differential privacy in a data-sharing system we need to decide where to store the data because it determines how we can add the noise. We can either store the data within each SME or gather it and store it somewhere else. A decentralized system would have a peer-to-peer architecture, where each SME would store, distribute and collect data from other SMEs. The centralized system would have a central control server or similar, that would be in charge of collecting and distributing the data.

Both architectures have their benefits and drawbacks, but there are a few special concerns when dealing with the type of data sharing systems in this project.
The first is that we are dealing with SMEs. As described earlier in section 3.1 SMEs are enterprises with at most 250 employees. These enterprises may have limited technical IT capabilities. A decentralized system would put more responsibility on the individual SMEs. Relying on each SME to be in charge of maintaining, operating, securing, etc. the system may be an unwanted and unreliable solution. Any system that would be implemented in the SMEs would put some amount of responsibility on the individual SMEs, but a centralized system moves some of that responsibility from the SMEs to the central operator.

This problem of distributed responsibility becomes greater when considering differential privacy. In the decentralized system, each SME's system would be in charge of data curation, introducing the right level of "noise" to maintain the desired level of privacy and accuracy. In the centralized architecture, the SMEs send their data to the central control server and this is where the server handles the curation of data based on the total data received from multiple SMEs. This can be implemented nationally, regionally, or in whatever group the system is implemented in.

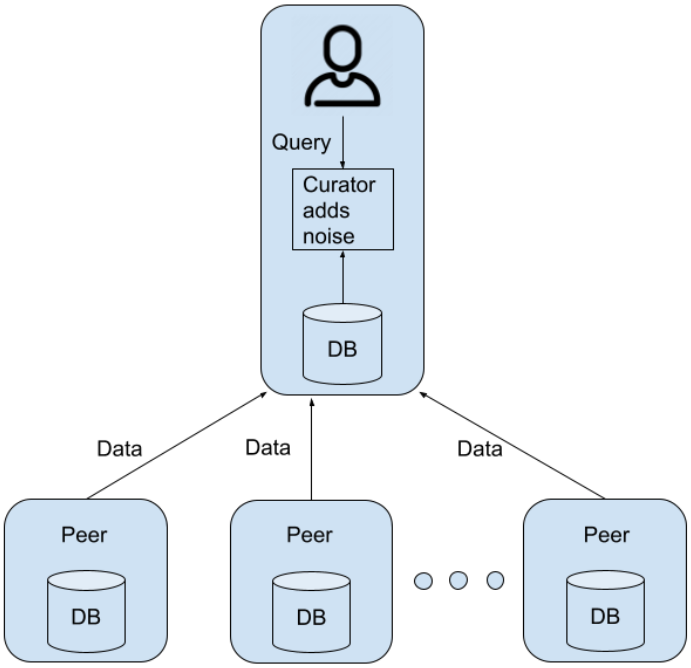
When we are using differential privacy, we maintain some level of privacy by adding noise. The goal of differential privacy is that the output of analysis on a given data set has a specified difference, ε, whether or not a single individual is included in the data set or not. The difference ε determines the level of privacy and accuracy. To achieve this, we look at the analysis function and determine its sensitivity to the removal or addition of a single individual. This gives us the amount of noise that needs to be introduced to the data [[27]]. This means that every query of data must include some level of noise. This gives us some options for centralization or decentralization.

One way of introducing noise into the system is to inject noise into each query. Each peer will perform the requested query on their data, inject the noise in the result and send it to the peer who made the request. However, in this model, the noise gets compounded as each query adds the required noise. However the noise can also be injected into the data set itself. This is done by adding noise to each data point in the data set.
While injecting noise into each data point distorts the individual values, quantities such as the mean are preserved as the random noise has an equal probability of being positive or negative, according to the laplace distribution. By injecting the noise into the data set we can send the entire data set with privacy protection. However we again end up with too much noise. This is because the amount of noise needed to achieve the desired level of protection becomes smaller as the size of the data set increases. These individual data sets will therefore require more noise than the combined data set.



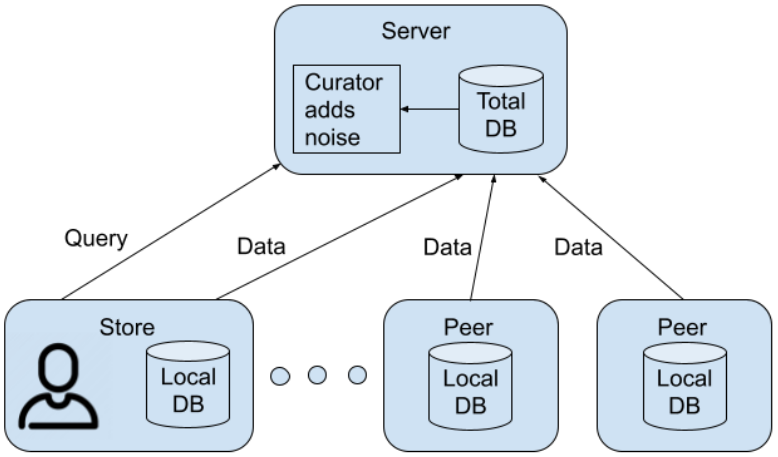
Peer-to-peer sender side noise injection.

To avoid this problem of compounding noise the peers can send the data sets without the noise, and then each peer continuously collects data, stores it without noise, and only introduces noise when the data set is queried.



Peer-to-peer receiver side noise injection.

This avoids excess noise in the data, but it requires every peer to store all the data without the protection of differential privacy. This has the downside of posing the risk that if any of the peer's systems are compromised all the data may be leaked without privacy protection. In our case the peers are SMEs and the goal is to have a large number of peers. These SMEs may not have the best IT security systems in place and so this may pose too large a risk. In the centralized architecture, we face similar problems.



Centralized architecture.

However, if the data is stored without noise in the central server and the noise is added when it is distributed then there is only one place where the data is stored without privacy protection, making it easier to secure.

One problem that the centralized architecture has is that it needs to be trusted. The SMEs need to trust the central data curator. If they do not trust the central curator they would be unlikely to use the system. Especially when their number one concern in data sharing is privacy, they would be unlikely to share data with an untrusted entity. Of course, we take into consideration laws and regulations such as GDPR that can be implemented throughout the European Union and the smaller groups within Europe. This can both incentivize SMEs to trust and exchange information with other enterprises using this centralized architecture. However, it can also discourage SMEs from participating in data sharing as we have mentioned previously (sec 3.7)

6.3 Problem of Repeat Queries

In differential privacy, the privacy protection degrades with repeated queries. Repeated queries allow the user to map the repeated query result to a probability distribution. As the query is repeated the user's probability distribution begins to be more accurate to the distribution that the added noise was taken from. If the user can perfectly map the probability distribution, then they can find the center of it, which is the true value of the query, before the added noise.
One way to try to solve this is to limit the number of times the same query can be performed on a given set of data. But as the focus of this project is on multiple SMEs sharing data and using the same system, this problem becomes worse. Limiting the number of times a query can be performed on a specific set of data, would usually alleviate the problem. However, in our case, the multiple organizations may share their query results, and thus reintroduce the problem.
This can be solved by only allowing a query to be performed once, globally, and then storing the result and providing it to other organizations that may make the same query. This solution becomes a little more difficult in a distributed system where the query results have to be disseminated throughout the system. In a centralized system, only the central server or central few servers needs to be kept up to date, while in a peer-to-peer system each peer needs to receive the same information.
Similarly, users may ask technically different queries that are functionally the same. For example, If a user asks for the number of enterprises that have more than 200 employees or if they ask for the number of enterprises that have more than 200 and less than 1 million employees. The answer to the question may be the same but the questions are technically different. Users could string together long sequences of conditions in their query, that make them technically different but give the same result.

This leads us to two possible models, the interactive and non-interactive. In the interactive model, the user of the model can perform any queries on the data set and do so adaptively, basing their next query on the previous one. The interactive model needs to be able to answer any query and any number of queries. This means that the accuracy of the answers must be low to maintain the same level of privacy. In the non-interactive model, all the queries are known beforehand. This means that the optimal level of accuracy, for the preferred level of privacy can be determined. The non-interactive model provides the most controlled level of privacy and accuracy, but it requires that all the queries that will be performed must be known beforehand. If we can develop a system where we know the queries that will be made then we can use the non-interactive model and provide a better balance of accuracy and privacy.

In our case, we are potentially dealing with a large number of SMEs and a large variety of different SMEs. Predicting all the possible queries that these SMEs may want to perform on the data, would be extremely difficult. Any query that is not accounted for in the non-interactive model would simply be unavailable, lowering the quality of the service. The best solution may be a mostly interactive system that tries to limit the number of repeated queries but still tries to provide the freedom necessary to handle the large variety of queries coming from the many SMEs.

Another potential solution is to design a non-interactive system for smaller groups of similar SMEs. By reducing the diversity in SMEs it will likely narrow down the number of different queries they wish to perform. This makes it easier to create a non-interactive system that still allows for all the desired queries. If any of the SMEs want to perform a new query that was not included in the non-interactive system, then the system can be updated to include it. This may change the specific amount of noise that needs to be added and may take some time to implement but it will allow new queries to be performed.

6.4 Problem of Bounding Sensitivity

One of the challenges of implementing differential privacy is determining the sensitivity of the query functions being performed on the data set. We have described the method for determining the sensitivity of a function in chapter 4.5 function sensitivity, but we also described the problem of bounding sensitivity. For the simple counting query, the sensitivity is always 1, as the output can at most be changed by 1 by a single addition, subtraction, or change to the data set. However, for other queries, the sensitivity can depend on the maximum value that an entry in the data set can have. For example, a summation query has sensitivity equal to the max amount an entry can have because the potentially added or subtracted entry could have the maximum possible value. This means we need to set a bound for the maximum values. When dealing with business data such as the data which is the focus of this project some data may have an easily defined bound such as product ratings.

Ratings often have a clearly defined set of possible values, defined during the development of the rating system. These values may be from 1-5, 1-10, and so on. These already specified bounds can easily and effectively be translated into the function sensitivity bound. However, some values do not have clear bounds, such as product sales or revenue.
If the bound is too low and an entry is added that is beyond the bound then the privacy protection will suffer. If the system is set up to provide a certain level of privacy protection, and an entry is added that lies beyond the function sensitivity bound, then the actual privacy protection provided will be lower than the intended level.

Clipping
A measure that can be used to alleviate this problem is clipping [30]]. When using clipping, any value that is beyond the bound is lowered to the value of the bound. This ensures that no value is beyond the bound, but this clipping may itself reveal information about the data, and clipping may also reduce the accuracy of the data set.

High Bound
Another measure is to set a very high bound. This solution reduces the risk that an entry falls outside the bound. The problem with this solution is that as the sensitivity bound increases the accuracy of the queries decreases. An unnecessarily high bound would result in an unnecessary low accuracy.

Categorizing Data
The last measure is the categorization of data. To find a good bound for the sensitivity, we need to understand what the data is. One way we can understand the maximum values for the data is to categorize the data source. In our project, we focus on SMEs. In the SMEs section of this report, we describe how the EU defines SMEs. The EU's goal is to define SMEs such that there is a better understanding and measures of SMEs. We can use these definitions to help us determine the bounds for some data that would be used in the system.

Company category	Staff headcount	Turnover	or	Balance sheet total
Medium-sized	< 250	≤ € 50 m		≤ € 43 m
Small	< 50	≤ € 10 m		≤ € 10 m
Micro	< 10	≤ € 2 m		≤ € 2 m

SME Definition [9]

These are the requirements for an enterprise to be defined as an SME by the EU. As we can see SMEs have been further categorized into Micro, Small, and Medium-sized enterprises. It may be possible to further break enterprises into smaller categories, however, there is a danger of excessive categorization. For example, if a new enterprise is added to the system then only the categories that the new enterprise falls within would change. If an adversary queries all the categories before and after and there's no other change, then the adversary would know which category the new enterprise falls within. And so the adversary would gain information based on the categories.

6.5 Problem of Choosing ε

As we have discussed in chapter 4.6 Privacy parameter ε, we use ε to change the probability of an adversary being able to correctly guess the data in the data set. This means we can freely choose the value of ε to fit whatever level of protection we need. However, there's a problem. A lower ε provides more protection but lower accuracy and thus lower utility as we described in chapter 6.1. So this is the tradeoff we need to make when we choose the value for ε.

One of the benefits of ε is that it can be chosen for each individual query. We don't need to set a global value for ε. We can set a value for each individual query based on how much protection the queried data requires.

Earlier in chapter 3.8 of this report, we looked at the potential consequences of the data protection being breached. We saw how the impact of a data breach can vary a lot, often depending on what data was breached. This allows us to look at the data and try to determine the potential consequences of the data being breached and set the value of ε accordingly. The problem here is that it can be very difficult to determine the consequences of a set of data being leaked. The possibility of de-anonymization attacks makes this even more difficult.

There's also the legal aspect. Regulations like the GDPR institute protection regulation on personal data and sensitive personal data [27], [31]. According to the GDPR, personal data is any data that is related to an identifiable person. Sensitive personal data includes Race and ethnic origin, Political beliefs, Religious or philosophical beliefs, Trade union affiliation, Genetic data, Biometric data for unique identification, Health information, Sexual relationships, or sexual orientation. From this, we know that we need to give extra protection, and thus a lower ε value to this type of information. The same goes for any data that requires special treatment under different regulations or laws.

The potential harm and the regulation surrounding the data can help developers determine what the value of ε should be. But ultimately there's no set value for ε, it is up to the developers to choose a value that the users of the system and the people that the data concerns are comfortable with, but still within the regulation and laws. What level of risk are people comfortable with? This will unfortunately likely vary from person to person and business to business. Trying to develop a data-sharing system where everyone is comfortable with the level of data risk may prove difficult. One way that it can be made easier is to implement the system in smaller groups where reaching a consensus is simpler.

As the system grows and more businesses join, the level of protection may need to be increased if more risk-averse businesses join. But here differential privacy has an advantage. As more data entries are added the lower the risk becomes. For example, if an adversary attempts to guess which business is paying their suppliers the most and there are only two businesses, the minimum probability of guessing correctly is 50%. However, if the number of businesses expands to 4 then the minimum is 25%. This means that as the number of SMEs join the system we may be able to lower ε and give more accuracy, without lowering the protection. This may be particularly important because, as we described earlier in chapter 3.1, SMEs make up 99% of all businesses in Europe.

This could provide a large number of businesses in the system. However, the first business to join would suffer either low protection or low accuracy until the system grows. However, if the system is implemented in smaller more specific groups, industry or community organizations could ensure that all or most of their members would join at the same time.

6.6 Resistance to Linkage Attacks

As we discussed in section 4.1, anonymization suffers from a weakness to linkage attacks. This weakness is one of the reasons we need to look for better privacy protection systems. So how does differential privacy fare against linkage attacks?

Linkage attacks work by finding links between the data set and some external data set. By finding matching data, the data can gain some level of certainty that the two entries in the data sets belong to the same person or entity. Here differential privacy can cause problems for the attacker by introducing noise in the data. By introducing noise to the data set the probability of finding matching data falls. The more noise is added the lower the probability of finding a match. Furthermore, the noise can cause two entries to start matching when in reality they concern two separate people or entities. This possibility of false links causes any links revealed by the linkage attack to be unreliable. Where simply anonymizing data fails to protect against linkage attacks, differential privacy both reduces the effectiveness of linkage attacks and causes their results to be unreliable. Differential privacy also does not impede the use of anonymization. It is still possible to anonymize the data set, by removing names and identifiers, and then use differential privacy.

6.7 Trust

One subject that has come up during the development of this project is trust in the system. Here we will briefly discuss this topic however we largely consider the solutions to these problems to be outside the scope of this project.
We need to consider the trust that the users of the data-sharing systems have in the system and the other participants. If the system is to provide a useful service then the users need to have some level of trust in the system and the other participants. The users need to be able to trust the system enough to share their data with the system, without fearing that the system or other entities will take advantage of or misuse the data. The users also need to trust the data they receive from the system, if they can't trust the data they receive then the data is useless.

If the users of the system misuse it and provide the other users with false or misleading data then the users that accept and rely on this data could be harmed by it. If a business relies on false data on user behavior or market trends, it may be led towards making bad business decisions. Any business using the system should look at the data received critically, however as stated, if the data is too unreliable then it loses its value. This means that the system would greatly benefit from some system or mechanism that prevents or discourages misuse of the system. Such a mechanism could attempt to verify trustworthy data, promote reliable data or discourage false data. Systems such as anomaly detection might be used. However, we consider the details and efficacy of these types of mechanisms to be outside the scope of this project.

As mentioned the users also need to be able to trust the system itself, and in a centralized model, the central server and data curator. The job of the data curator is to use the tools of differential privacy to achieve the desired level of data privacy when the data is queried. Thus the users need to be able to trust that the curator is implemented properly and is applying the right desired level of privacy.

The curator also has access to the data without the privacy protection, as it is the curator that applies it, so the users need to be able to trust that the curator is not taking advantage of the unprotected data. If the system implements a centralized control server, then this server needs to be placed under the control of a trustworthy organization.

We need the organization to not be incentivized to take advantage of the data or to treat the users differently, thus creating an unfair system. If we place the system in the care of a business, then there would be an incentive for them to take advantage of the system and the data, because they may be able to profit from doing so.

This means we need the system to be controlled by an organization that is not interested in taking advantage of the system to gain profit, but instead an organization whose goals align with the goals of the system, namely the sharing of information between businesses. One such organization could be the EU. In an earlier chapter 3.2 on the EU and data sharing, we describe how the EU is trying to implement regulations that will make sharing data between businesses easier. From this, we can see that the EU may be interested in the system and that its goals align with that of the system. This makes choosing the EU to be in charge of the system a good option. Similarly, there are many industry or community organizations whose interests are aligned with the businesses of their community or industry. These organizations may already have some level of trust from the SMEs. This would make these organizations a good choice for the SMEs in those specific industries or communities.

6.8 Differential Privacy libraries

Differential privacy is not just a theoretical method for privacy protection. A number of libraries have been developed to allow businesses to implement differential privacy. Among these libraries are:
IBMs, Diffprivlib v0.5 for python[32]
Benjamin I. P. Rubinstein and Francesco Aldàs, diffpriv for R [33].
Google's DP building block libraries for C++, GO and Java [34].
OpenDPs, OpenDP, for python [35].

As we can see there's a wide interest in developing differential privacy libraries. These differential privacy libraries are developed by individual research projects, research communities, and large organizations such as Google and IBM. These are all open course projects that allow people and businesses to view, contribute, and/or implement these libraries.
These libraries have many of the features of differential privacy but many are also under development. We can see from Google's DP building block libraries that some features are still in development.

Algorithm	C++	Go	Java
Laplace mechanism	Supported	Supported	Supported
Gaussian mechanism	Supported	Supported	Supported
Count	Supported	Supported	Supported
Sum	Supported	Supported	Supported
Mean	Supported	Supported	Supported
Variance	Supported	Supported	Supported
Standard deviation	Supported	Supported	Planned
Quantiles	Supported	Supported	Supported
Automatic bounds approximation	Supported	Planned	Supported
Truncated geometric thresholding	Supported	Supported	Supported
Laplace thresholding	Supported	Supported	Supported
Gaussian thresholding	Planned	Supported	Supported

DP building block libraries supported algorithms[34]

Likewise, the OpenDP states: “OpenDP is under development, and we expect to release new versions frequently, incorporating feedback and code contributions from the OpenDP Community.”[35].

While these libraries are still in development we can see that they have developed many of the main functionalities of Differential Privacy we have discussed throughout this report. Our research on these libraries has also shown that there are difficulties in implementing these libraries. However, these libraries are almost fully developed and once the final features have been developed and the implementation difficulties have been smoothed out, these libraries could function well for data sharing purposes.

6.9 Technology Readiness Level

The method is used here to illustrate the journey of differential privacy throughout the years. Furthermore it enhances our project as it showcases some key elements throughout its history that can illustrate whether or not, differential privacy can act as a suitable solution to our problem formulation.

The technology readiness level is a method first introduced by NASA in the 1970s for the estimation of the development of technology from the acquisition of the said technology until the end of the program. It has nine total levels, each level showcases the maturity of the technology.

The concept of differential privacy has been around since 2006 with some basic principles dating back to the 1970s. Through the years countless reports have been made to showcase the technology and some companies have even managed to implement differential privacy methods on some of their services. However, differential privacy has not seen commercial implementation so far[36].

Research	**1. Basic principles Observed** **2. Technology Concept Formulated** **3. Experimental Proof of Concept**	Basic principles of this technology can be tracked all the way back from the 70s when mathematics was used to impose privacy properties on statistical databases [37]. The main concept came into tuition in 2003. The concept of noise was introduced in order to emphasize that a database can not truly withhold private information when it is published without the addition of noise [38]. In 2006 further noise calibrations were used to enhance the functionality of the information [39].
Development	**4. Technology Validated in Lab** **5. Technology Validated in Relevant Environment** **6. Technology Demonstrated in Relevant Environment**	Concepts such as ε-differential privacy were also introduced. Different organizations have implemented various forms of differential privacy. In 2008 a paper used differential privacy on U.S. Census Bureau commuting patterns [40]. In 2015 Google used differential privacy to track and share traffic information [41]. Microsoft used it in 2017 to collect telemetry data [42]. The list of companies and organizations that have used concepts of differential privacy for specific purposes goes on.
Deployment	**7. System Prototype Demonstration in Operational Environment** **8. System Complete and Qualified** **9. Actual System Proven in Operational Environment**	Despite all the solutions that differential privacy can bring to the table, an actual implementation for commercial purposes has not been implemented yet. There are several papers and specialized implementations by tech giants such as Apple and Google but the reality is that it has not yet achieved commercial success. One factor that can be responsible for this is that differential privacy is not a technique but a set of mathematical theories on what privacy is and how in theory it could be achieved [43].

7 Discussion

The analysis section of this report has presented a lot of topics that need to be discussed when considering how differential privacy can be used for data sharing between SMEs.

Our first takeaway from our research on SMEs is that it is the minority of SMEs that share data for the purposes of sharing and gathering information. However, the proportion may be growing. The fact that it is a minority of SMEs that share data means that for many SMEs the prospect of using differential privacy is not just competing with other technologies, but it needs to make data sharing as a whole good enough for them to participate. For those SMEs that are already sharing data, the question is whether differential privacy is better than their current method. But for those that don't already share data, the question is whether the use of differential privacy will make data sharing worth it in the first place. This means that the solution needs to require even less of the SMEs. Starting to share data already creates an administrative burden on the SMEs. If the privacy protection solution requires a large amount of resources to use, then the total burden may be too large for the SMEs to handle.

7.1 K-anonymity vs Differential Privacy

A discussion must be made on the various elements that surround the application of k-anonymity and that of differential privacy. k-anonymity as we have mentioned throughout this report is a way of publishing data structures or tables while protecting personal and sensitive information. The publication of such data tables, however, is met with the choice between either making the information of the data structure more impactful at the risk of loss of disclosure or making the data set more secure at the risk of limited information effectiveness.

k-anonymity uses generalization and suppression methods to ensure that sensitive information can not be linked. There are some elements that these methods must consider when they are implemented. The first is the Identifiers, these may be in the data structures and consist of information that can identify individuals such as passport numbers or social security numbers. Another element is the key attributes which are information that can be used in combination with external information in order to identify individuals from the published database. Lastly, we have the confidential outcome attributes that consist of sensitive information such as salary, health conditions, or religion. Generalization and suppression methods must comply with the definition of k-anonymity. "A protected data set is said to satisfy k-anonymity for k > 1 if, for each combination of key attributes, at least k records exist in the data set sharing that combination" [44].

k-anonymity is a more simplistic concept in terms of implementation than that of differential privacy. This is because the latter is not a method or an algorithm but a mathematical term. There are some drawbacks however to its simplicity. As we have mentioned throughout the report k-anonymity is prone to several and different in nature attacks that can be used to identify and retrieve critical or sensitive information from data tables. Even after privacy enhancing techniques such as l-diversity and t-closeness, k-anonymity can still be attacked by skewness attacks and similarity attacks [44].

With its simplicity, it also offers a limited set of controls over the privacy to usefulness of data ratio. Since it relies on its two main ways of inflicting privacy, it can not give as many options as differential privacy. An assumption that can be made however is that SMEs might have an easier time understanding and implementing k-anonymity in comparison with differential privacy, however, the lack of solid protection of privacy might discourage them in the end.

The big benefit of differential privacy is that it allows the user to control the tradeoff between privacy protection and utility. This means that the users of the system can decide on the level of protection and utility that they are comfortable with. The users can tailor the level of protection to their situation and to the data that is being shared. This makes differential privacy very flexible and allows it to be used in many circumstances. However, differential privacy does have its limits. Some SMEs may find that it is impossible to find a level of protection that also provides enough utility for data sharing to be useful. This will highly depend on how sensitive the data is and how risk-averse the SME is.

The tradeoff between privacy and utility also means that implementing differential privacy will make data sharing less impactful as noise is added to the data obscuring it. A market that uses differentially private data sharing will not be fully transparent. But the market will be significantly more transparent than without data sharing.

Differential privacy does have some issues, such as the problem of repeat queries and bounding function sensitivity. These problems require careful consideration when the system is implemented but they are not insurmountable. We have described some tools that can be used to work around the problem of repeat queries such as the interactive and non-interactive models.

We have also described some methods that can be used to help find good bounds for function sensitivity. These tools and methods allow developers to work around these problems. Differential privacy also works well against the type of privacy attacks that anonymization fails to protect against. By introducing noise differential privacy effectively protects against attacks such as linkage attacks. Differential privacy can also work in combination with anonymization, and can therefore be a useful tool in addition to other privacy protection methods.

Overall differential privacy provides a lot of flexibility and protection and the problems it does have can be worked around. The main issues lie in implementing differential privacy. When considering the implementation of differential privacy in data sharing we first need to look at the system architecture, because it has a big effect on how differential privacy is implemented.

7.2 Architectures

In a peer-to-peer architecture- the noise can either be added on the senders' side or the side of the receiver. If it is implemented on the sender's side we face the problem of compounding noise. If it is implemented on the receiver's side, then the sender needs to trust the receiver with their unprotected data. Since privacy is the SMEs' primary concern in data sharing, the second option seems unlikely to be their preferred option. Additionally, according to the GDPR, the peers would need to form contracts with each other peer, before they can process the data.

Another architecture is a centralized architecture where noise is added not at the sender or receiver but at some other server where the data is gathered and distributed. In this model, we avoid the compounding noise and avoid requiring SMEs to fully trust each other with their data. It does introduce the problem that the SMEs need to trust the central server and the server needs to be able to handle the potentially massive amount of data. In the centralized architecture, the server is processing the data and so the SMEs need to form a contract with its controller.

This is where the discussion of the scale of the system comes in. In any architecture, the SMEs need to agree on the parameters of the system as well as each other or the central server. In this report, we have focused on SMEs in the EU as the EU has an interest in data sharing and SMEs. However, creating a single centralized system for the entire EU appears to be infeasible. Aside from the technical problems of handling such a large amount of data and traffic, there are also problems specific to differential privacy. The main benefit of differential privacy is the ability of the users who are sharing data to control the tradeoff between privacy and utility. However, this means that all the SMEs that are sharing data need to agree on what that tradeoff should be. In a massive system, such as one that covers the entirety of the EU, it is likely impossible to find a tradeoff that every SME is happy with.

As we have just described, differential privacy can be very intricate and complicated. There's a lot of ways that differential privacy can be implemented and configured. This can be a big benefit, but it also has a drawback. In our research on data sharing in SMEs we found that some SMEs don't have any experience with data sharing. One of our interviewees complained of a lack of digitalisation and another said that the industry they were in was slow to innovate. Some of our interviewees also said that they didn't have the resources to dedicate to data sharing. A system like differential privacy that may be complicated to implement and configure, could be too big a burden for these SMEs to carry. If differential privacy was simpler to implement then it may be more useful to them. This is where grouping of SMEs may play a role. If an industry organization were to implement a data sharing system using differential privacy, they could take some of the workload from the SMEs. This would allow the SMEs to use a differentially private data sharing system without having the resources and sufficient digitalisation to run the whole system on their own.

A data-sharing system does not need to cover the entire EU. One of our findings from our research is that SMEs tend to gather data from their supply chain partners and their local competitors. Smaller groupings of SMEs could join together and establish their own differentially private data sharing system.

Smaller groups of more similar SMEs would perhaps be more likely to agree on the set of parameters required for differential privacy. Existing industry or community organizations that already have a number of SMEs as members could more easily implement a centralized system controlled by the organization.

The scale and focus of the system has a big impact on the utility and feasibility of a differentially private data-sharing system. With a focused group of SMEs, a non-interactive model for querying would be easier to implement and the bound of the function's sensitivities would be tighter, both of which would provide more accuracy or protection. However, the protection and accuracy provided by differential privacy also scales with the amount of data that is being queried. The ideal system would be a system with large amounts of data, but where everyone agrees on all the parameters. However, this ideal situation is probably unlikely.

The focus of the data sharing is also important as to whether differential privacy would work at all. Differential privacy is only able to handle numerical data. This is because of the way it adds noise. Adding noise this way does not work for qualitative data, such as text.
We found in our interviews that SMEs share both numerical and non-numerical data. SMEs share numerical data such as prices and wages, but also non-numerical data such as ideas and processes. This means that Differential privacy is only useful for part of the type of data that SMEs share.

However, there may also be some types of data sharing where the SMEs do not want privacy protection at all. In some cases, they wish to share data that includes private information. For example, we interviewed some SMEs that use an external booking system. These booking systems are run by another business. In this case, the data shared would need to include information about the customer and the booking. This data should not be anonymized or have noise introduced, because they need the specific information, and do not need to hide it.

Even when the data does not need to be unobfuscated, Differential Privacy's focus on unlinking the data from the data provider may be a detriment. We found that one of the main benefits that SMEs gain from data sharing is the strengthening of business relations. If the shared dataset is entirely unlinked from the data provider then that hinders their ability to build business relations.

However, if the data contributors are identified, as part of the metadata, but the individual data points cannot be linked to any one of the contributors, then it is possible to build business relations while maintaining privacy. But building business relations may be difficult this way as it is not possible to see what data the other business contributed.

Privacy concerns appeared to be the main barrier to data sharing in our research into data sharing in SMEs. As we have seen in our examination of differential privacy, differential privacy is able to provide some level of privacy protection. Differential privacy in combination with anonymization is able to provide SMEs with a level of privacy protection that they can tailor to their needs. This should remove or significantly reduce this barrier to data sharing.

However, as we have described, differential privacy is already in use in some places and it is not the only tool that can provide privacy protection. K-anonymity is also able to provide some level of privacy protection and is also seeing use. This may suggest that there are other barriers to data sharing.

The last part of our interviews with SMEs was a question about differential privacy or similar tools. We described a tool that would allow SMEs to share data without sharing specific details and without it being linked to them. We then asked if they would be willing to share data using such a system. 10 out of the 13 SMEs would be willing to use such a system. However, some did have caveats such as, they would use it if they saw a benefit from using it, or if the business was bigger.

The fact that differential privacy is in use in some places and the fact that the SMEs we interviewed would be willing to use a system that we described to them suggests that it is not some failing of the concept of differential privacy itself that is the reason that it is not being used. What may be the real barrier to the adoption of differential privacy in data sharing among SMEs is a lack of ready-to-use tools and/or a lack of awareness.

One of the barriers to data sharing we found from our research was the lack of resources to dedicate to data sharing. It was some of the SMEs' perception that they did not have the time and/or manpower to dedicate to data sharing. As we have seen from our analysis of the technology readiness level of Differential Privacy, data sharing tools using differential privacy are not commercially available. There are publicly available libraries for Differential Privacy, but some are still in development. SMEs may also find it to be too difficult or costly to develop their own systems using these libraries. If Differential Privacy is to be used in data sharing among SMEs, it would seem that fully developed systems and tools that are easy to use and implement would need to be available.

Our research suggests that privacy concerns are the main barrier to data sharing. However, a large portion of the SMEs we interviewed claimed that they would share data if they were able to share it without the data being linked to them. It appears that privacy protection may not be the main barrier to data sharing, but instead the main barrier is that Differential Privacy or other similar systems are not developed and available enough for SMEs.

7.3 Recommendations

We have analyzed many aspects of differential privacy in data sharing between SMEs, and after discussing these topics we have found that a full data sharing system is needed for the SMEs to start sharing data. This brings up the question of how such a system should be designed to best facilitate data sharing between SMEs and to best meet the SMEs’ needs.

We believe that differential privacy could be very useful in a centralized data-sharing system between groups of SMEs. Utilizing existing industry or community organizations allows differential privacy to be more easily implemented and it could provide the SMEs with valuable data sharing between similar SMEs in the organizations.

By implementing a non-interactive and centralized system, differential privacy could offer high degrees of utility and privacy. The system needs to be fully developed and as ready to use as possible to allow for easy implementation and use within the SMEs.

In larger scopes, a peer-to-peer system with an interactive model would provide larger amounts of freedom and avoid the feasibility problems of a centralized system. This would in turn require more protection but it would benefit from the larger amounts of data.
The down side to such a system would be the larger investment of the SMEs resources to implement and use. The peer-to-peer system may be better suited for larger businesses with more resources to invest.

While both architectures seem to be suitable solutions, the problem remains with the feasibility to understand differential privacy as a whole. On top of what we have discussed throughout the report which includes the positives and drawbacks of differential privacy and how the different configurations might impact its overall performance. This does not mean however that it necessarily bodes well with the idea to commercialize this mathematical theory.

7.4 Future Work

In this project, we dove into how data sharing principles with the implementation of differential privacy can assist and boost the overall information of SMEs. An analysis has been carried out throughout the report, stating how impactful differential privacy is and what positives and drawbacks surround this technology. This whole project remains a theoretical approach to what differential privacy can bring to the table, however, an implementation can be the future work of this project. By implementation, we mean an actual system that can engage SMEs to exchange information safely with the assistance of differential privacy. However several factors must be examined.

Research must be carried out to examine the specific needs of SMEs. Each SME may have a unique need or desire for a specific configuration of differential privacy. We need to consider if they are willing to engage with our system, and at what rate. How much noise they want and how much transparency. Moving onwards, adjustments to the system must be made depending on the scale of it. It can be applied nationally, regionally, or depending on the sector of each enterprise.

All in all, the acquisition of data, relevant to the decision-making, when designing the system is vital for future work. It will be used to create a prototype of the system. In time the system will be tested and modified to find the best possible solution for each grouping of SMEs.

8 Conclusion

In order to conclude this report, we must examine if we have answered the questions and the problem formulation stated at the beginning of our paper. The problem formulation along with the sub-questions have guided the overall research and decision-making of this report. We start things off by answering the sub-questions.

- **How crucial is the need for small to medium-sized enterprises to protect their privacy?**

Throughout the report, we have established that SMEs benefit from the exchange of data with other enterprises. However, the lack of information within SMEs comes from the fact that they are hesitant to exchange information. Privacy concerns and legal technicalities have impeded SMEs from data sharing. We have also shown the potentially great harm that can be caused to both the SMEs and their customers in case of privacy protection failure. Therefore the protection of privacy becomes a crucial matter for the SMEs both in terms of protection and in terms of legality.

- **What are the problems of implementing differential privacy in data sharing?**

To start things off we must admit that an actual implementation of a system that uses differential privacy to protect the exchange of data is out of the scope of this project.
This needs to be stated as an actual implementation might procure further technicalities and additional drawbacks. For differential privacy one of the main benefits has also proven to be one of the main drawbacks. The adjustment of noise in the data can be beneficial but at the same time a problem for some SMEs. It is mentioned throughout the report that too much noise will deteriorate the quality of the information that is exchanged. At the same time, too little noise can be harmful to the protection of private information. By default, even the slightest amount of noise limits the transparency of data and therefore the overall transparency of the market as a whole. Technical issues such as repeated queries and bounding function sensitivity can also be problematic during implementation. Along with these problems, there can be a debate on the scale of the system and on the architecture. Defining all these parameters during implementation so that they are satisfactory to all parties could prove very difficult.

- **What benefits can differentially private data sharing provide?**

Differential privacy enables SMEs to conduct data sharing in a privacy preserving way. This alone allows enterprises to expand their pool of information and in turn become more competitive. It allows them to improve relations and collaborations with their stakeholders, deliver a better product, and even achieve legal compliance. Just as we mentioned in the previous paragraph the level of noise can be adjusted; this allows SMEs to be in charge of how transparent their data sharing will be. The scaling of the system can also be adjusted. It can be implemented on a national level or depending on the enterprise’s market sector. Furthermore, the more enterprises enter the system the more effective differential privacy becomes. Protection and privacy scale according to the amount of information.

Now that we have answered our sub questions it is time to finally gather our conclusions and answer our problem formulation.

“Is differential privacy suitable for protecting privacy and confidentiality in data sharing among small to medium-sized enterprises?”

While differential privacy can be suitable, it is not currently at a stage where it is suitable for small to medium sized enterprises. As we have written in previous sections, it has both benefits and drawbacks. Differential privacy can provide flexible protection in data sharing, where SMEs are capable of adjusting the parameters to suit their needs. Differential Privacy’s complex mechanics allows for a greater degree of control and flexibility than the alternative method of k-anonymity. We have shown that it can work in different configurations and at different scales. However we believe that smaller local configurations are more suitable to the SMEs.

Nevertheless, implementation could be problematic as reaching a consensus on the configuration may prove challenging. The problems in implementations are especially problematic as we have found that some SMEs don't have a lot of resources to dedicate to data sharing, and so a fully developed and commercially ready system would likely be necessary before SMEs would be willing to adopt it. But such a system is not currently available. Differential privacy appears to be a promising method for improving data sharing, but it does not currently appear to be mature enough to be suitable for small to medium sized enterprises.

9 Reference List

[1] MicroStrategy, “2020-Global-State-of-Enterprise-Analytics.” 2020 [Online]. Available: https://www3.microstrategy.com/getmedia/db67a6c7-0bc5-41fa-82a9-bb14ec6868d6/2020-Global-State-of-Enterprise-Analytics.pdf

[2] corporate-body. CNECT:Directorate-General for Communications Networks, Content, Technology, and N. A. E. Benelux, *Study on data sharing between companies in Europe : final report*. Publications Office of the European Union, 2018.

[3] “New survey finds British businesses are reluctant to proactively share data.” [Online]. Available: https://theodi.org/article/new-survey-finds-just-27-of-british-businesses-are-sharing-data/. [Accessed: May 31, 2022]

[4] Ninghui Li Purdue University, W. Lafayette, IN, and USA, “Differential Privacy in the Local Setting,” *ACM Conferences*. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3180445.3190667. [Accessed: May 31, 2022]

[5] G. Winter, “A Comparative Discussion of the Notion of ‘Validity’ in Qualitative and Quantitative Research,” *The Qualitative Report*, Mar. 2000 [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.533.3639&rep=rep1&type=pdf. [Accessed: Sep. 25, 2022]

[6] E. Knott, A. H. Rao, K. Summers, and C. Teeger, “Interviews in the social sciences,” *Nature Reviews Methods Primers*, vol. 2, no. 1, pp. 1–15, Sep. 2022.

[7] “Diffusion of Innovation Theory.” [Online]. Available: https://cjni.net/journal/?p=1444. [Accessed: Sep. 27, 2022]

[8] C. Shapiro, Shapiro, H. R. Varian, and S. Carl, *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Press, 1999.

[9] “SME definition,” *Internal Market, Industry, Entrepreneurship and SMEs*. [Online]. Available: https://ec.europa.eu/growth/smes/sme-definition_en. [Accessed: May 31, 2022]

[10] “SMEs,” *Internal Market, Industry, Entrepreneurship and SMEs*. [Online]. Available: https://ec.europa.eu/growth/smes_en. [Accessed: May 31, 2022]

[11] European Commission, “Synopsis report of the public consultation on building a European data economy,” *Shaping Europe’s digital future*, Sep. 07, 2017. [Online]. Available: https://digital-strategy.ec.europa.eu/en/synopsis-report-public-consultation-building-european-data-economy

[12] PWC, “Data exchange as a first step towards data economy,” 2018 [Online]. Available: https://internationaldataspaces.org//wp-content/uploads/Data-exchange-PWC-study.pdf. [Accessed: Sep. 04, 2022]

[13] A. Hayes, “Perfect Competition,” *Investopedia*, Nov. 24, 2003. [Online]. Available: https://www.investopedia.com/terms/p/perfectcompetition.asp. [Accessed: May 31, 2022]

[14] P. Swabey, “Policymakers want businesses to share more data. How might it work?,” *Tech Monitor*, Sep. 27, 2021. [Online]. Available: https://techmonitor.ai/policy/digital-economy/policymakers-want-businesses-to-share-more-data-how-might-it-work. [Accessed: May 31, 2022]

[15] “The Value of Data - Summary report 2020,” *Bennett Institute for Public Policy*, Feb. 25, 2020. [Online]. Available: https://www.bennettinstitute.cam.ac.uk/publications/value-data-summary-report/. [Accessed: May 31, 2022]

[16] “Data protection under GDPR,” *Your Europe*, Mar. 23, 2018. [Online]. Available: https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_en.htm. [Accessed: May 31, 2022]
[17] S. J. De and D. L. Métayer, “Privacy Risk Analysis,” *Synth. Lect. Inf. Secur. Priv. Trust*, vol. 81, no. 3, pp. 1–133, Sep. 2016.
[18] “Fines / Penalties - General Data Protection Regulation (GDPR),” *General Data Protection Regulation (GDPR)*. [Online]. Available: https://gdpr-info.eu/issues/fines-penalties/. [Accessed: May 31, 2022]
[19] P. Ohm, “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,” *SSRN*, Jul. 13, 2012. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006. [Accessed: May 31, 2022]
[20] A. Narayanan and V. Shmatikov, “Robust De-anonymization of Large Sparse Datasets,” *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008 [Online]. Available: http://dx.doi.org/10.1109/sp.2008.33
[21] K. Hafner, “And if You Liked the Movie, a Netflix Contest May Reward You Handsomely,” *The New York Times*, The New York Times, Oct. 02, 2006 [Online]. Available: https://www.nytimes.com/2006/10/02/technology/02netflix.html. [Accessed: May 31, 2022]
[22] “IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows,” *IMDb*. [Online]. Available: https://www.imdb.com/. [Accessed: May 31, 2022]
[23] P. Samarati, “Protecting Respondents’ Identities in Microdata Release,” Nov. 01, 2001. [Online]. Available: https://spdp.di.unimi.it/papers/tkde_k-anonymity.pdf. [Accessed: Sep. 04, 2022]
[24] “Tuple.” [Online]. Available: http://encyclopediaofmath.org/index.php?title=Tuple&oldid=11398. [Accessed: Sep. 04, 2022]
[25] L. Sweeney, “k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, May 2002. [Online]. Available: https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf
[26] P. S. Y. Charu C. Aggarwal, “A GENERAL SURVEY OF PRIVACY-PRESERVING DATA MINING MODELS AND ALGORITHMS,” *Privacy-Preserving Data Mining. Advances in Database Systems, vol 34. Springer, Boston*, 2008. [Online]. Available: http://charuaggarwal.net/generalsurvey.pdf
[27] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2013.
[28] C. Halton, “Diffusion of Innovations Theory,” *Investopedia*, Jan. 17, 2011. [Online]. Available: https://www.investopedia.com/terms/d/diffusion-of-innovations-theory.asp. [Accessed: May 31, 2022]
[29] “Differential Privacy.” [Online]. Available: https://privacytools.seas.harvard.edu/differential-privacy. [Accessed: May 31, 2022]
[30] “Sensitivity — Programming Differential Privacy.” [Online]. Available: https://programming-dp.com/notebooks/ch5.html. [Accessed: May 31, 2022]
[31] “General Data Protection Regulation (GDPR) – Official Legal Text,” *General Data Protection Regulation (GDPR)*. [Online]. Available: https://gdpr-info.eu/. [Accessed: May 31, 2022]
[32] “GitHub - IBM/differential-privacy-library: Diffprivlib: The IBM Differential Privacy Library,” *GitHub*. [Online]. Available: https://github.com/IBM/differential-privacy-library. [Accessed: Sep. 04, 2022]
[33] Benjamin I. P. Rubinstein & Francesco Aldà., “Pain-Free Random Differential Privacy with Sensitivity Sampling,” *GitHub*, 2017. [Online]. Available: https://github.com/brubinstein/diffpriv
[34] “GitHub - google/differential-privacy: Google’s differential privacy libraries,” *GitHub*. [Online]. Available: https://github.com/google/differential-privacy. [Accessed: Sep. 04, 2022]

[35] “OpenDP,” *GitHub*. [Online]. Available: https://github.com/opendp. [Accessed: Sep. 04, 2022]
[36] M. Héder, “From NASA to EU: the evolution of the TRL scale in Public Sector Innovation,” Aug. 2017 [Online]. Available: https://www.innovation.cc/discussion-papers/2017_22_2_3_heder_nasa-to-eu-trl-scale.pdf. [Accessed: Sep. 04, 2022]
[37] T. Dalenius, “Towards a methodology for statistical disclosure control,” 1977 [Online]. Available: https://ecommons.cornell.edu/handle/1813/111303. [Accessed: Sep. 04, 2022]
[38] I. Dinur and K. Nissim, “Revealing information while preserving privacy,” in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS ’03*, San Diego, California, 2003, doi: 10.1145/773153.773173 [Online]. Available: http://portal.acm.org/citation.cfm?doid=773153.773173
[39] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” *Theory of Cryptography*, pp. 265–284, 2006.
[40] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, “Privacy: Theory meets practice on the map,” in *2008 IEEE 24th International Conference on Data Engineering*, Cancun, Mexico, Apr. 2008, doi: 10.1109/icde.2008.4497436 [Online]. Available: http://ieeexplore.ieee.org/document/4497436/
[41] “Tackling Urban Mobility with Technology,” *Google Europe Blog*. [Online]. Available: https://europe.googleblog.com/2015/11/tackling-urban-mobility-with-technology.html. [Accessed: Sep. 04, 2022]
[42] B. Ding, J. (jana) Kulkarni, and S. Yekhanin, “Collecting Telemetry Data Privately,” in *Advances in Neural Information Processing Systems 30*, Dec. 2017 [Online]. Available: https://www.microsoft.com/en-us/research/publication/collecting-telemetry-data-privately/. [Accessed: Sep. 04, 2022]
[43] I. Nerurkar, “Differential Privacy from theory to practice,” *LeapYear*, Feb. 07, 2020. [Online]. Available: https://leapyear.io/resources/blog-posts/differential-privacy-from-theory-to-practice/. [Accessed: Sep. 04, 2022]
[44] J. Domingo-Ferrer and V. Torra, “A critique of k-anonymity and some of its enhancements,” in *2008 Third International Conference on Availability, Reliability and Security*, Mar. 2008, doi: 10.1109/ares.2008.97 [Online]. Available: http://ieeexplore.ieee.org/document/4529451/

10 Appendix

10.1 Interview transcripts

Interview 1

What kind of a SME do you own ?

- It is a tourist shop that sells handmade greek products.

Are you familiar with data sharing?

- Not particularly, the only thing attached to that is maybe sharing information through social media

So you participate in data sharing through social media, how do you exactly do that ?

- I would promote certain items in the shop that i want to sell, so i way for me to come to a wider audience is to put it upon my particular page and try to push the sale through there i would also get ideas for other items that i like, maybe to order or by looking through other businesses with similar products to myself.

When do you usually do that and are the other businesses similar to yours or are you also looking at different establishments?

- Different also, ideas for everything, it could be the decor or something i want to change or new products that i would think to sell in my shop that I do not supply at the moment or look at the setup that other businesses have. It is giving me a lot of ideas and I use it a lot throughout the day. Whenever I am not busy with my shop, I will be on my computer doing this.

You are getting some benefits out of it, do you think you could be getting more out of it?

- If I knew how to use it better then I think I would get more out of it.

How do you think you could use it better?

- Ok I do not have a great knowledge of it, I presume I could use it better to sell more products. I don't know in which other way I could use data sharing better.

If you are not certain as to how then what would you like to know more of like I can give you an example, wages, equipment, service, marketing, advertisement other trade secrets?

- Probably a little bit of everything. Okay wages do not affect me because I am the sole owner and I don't have any staff. Marketing for sure, advertising is what I am doing through my page. Improving, being more efficient.

Now from your position what type of data would you be willing to share with the public?

- Like?

It could be through your shop the products that you sell, the location or trade secrets with other SMEs

- Yes I could do that also I would not have a problem with that. If there was a way that I could promote another company’s product or pass that on to a similar business that is located obviously in the same region with me.

Is there anything you are worried about when doing so ?

- Probably safety, you know if someone would be able to hack into or something like this but other than that..

So privacy concerns?

- Yes privacy concerns

Do you know or do you see any problems in information sharing?

- No, it is something that I have probably not thought a lot about, so no I do not think that there would be any problems if it was used correctly.

Would you be willing to participate in a system where you can share private data without the specific details and without it being linked to you?

- That is a tricky question, I am not sure.. Anything that we hear about sharing private details or whatever I think puts us a little bit on guard today in the system or what it is. It would make me a bit weary of doing it but not that I wouldn't be willing to.

Interview 2

What kind of a store do you own?

- I own a bookstore that specializes also on printing papers, posters and more.

Does information sharing mean anything to you?

- Well yes ok, because people use the internet a lot these days. Especially now because people use their smartphones to exchange information. They send me files and pictures for me to print or to scan in order to send them to some enterprise. It is a way of life now, this constant exchange of information.

Do you participate in information sharing?

- Yes because we also use it as an enterprise. However we do not exchange so much as to say personal information. It is mostly information that we use on a daily basis, something that we already know. For example we do not use any personal information in order to do business with the customers but we act as the middle man as they send us some documents in order for us to print them or to scan them which may contain their personal information. However, we are not interested in that.

Do you participate in other information sharing such as exchanging data about wages, equipment or trade secrets in order for you to maybe offer a better product?

- Yes because when we are interested in something we go online and we see products that we are interested in. Nowadays we communicate for example with suppliers through email discussing the selection and volume of products that we need. The old ways of looking at flyers or through a magazine that specializes in bookstore products has now been lost.

If you do participate in information sharing, what are the benefits of it?

- For us actually there aren't any benefits, in fact it is a burden. This is because we offer a service which we are not getting any money or credit for. For example someone comes inside the store saying that they have a file which they do not know how to send or print and they ask for help. Of course I want to give my services to the customer so I will help them figure it out. The 5-10 or even 15 minutes that it takes for me to figure out what the customer wants to send, scan or print is time that I am not getting paid for. I will only get paid for printing their file but not for showing them how to send the file. Printing a page or two costs like 50 cents which is nothing. Imagine this happens multiple times on a daily basis. So you spend hours trying to help customers with their devices on how to access their files for you to get paid five euros in total.

Do you think that you could be getting more out of this?

- No because essentially we act as a middle man for the state. We have essentially taken the burden that the country has imposed. For example if someone goes to a state agency to order some affairs the state agency might ask them to send them a file through email. The person that does not know how to do that will figure that the best thing to do is to go to a bookstore in order to send the file to that state agency. Even for example electricity bills during covid times half of the people on this island did not know how to pay their bills online. We got the burden of the state because the

state decided to become digitized without the people necessarily knowing how to perform these online actions.

What kind of data would you like to share and what kind of data would you like to receive?

- I think that we are going through a phase where the state says that we should become more digital but the people of this country are not ready yet for this step. I say this because we are in a hybrid phase. It is not enough to have all your information digitized on your smartphone for example. Because state agencies ask for physical papers and digitized files at the same time. The digital part is still young in this country and until now we have done everything as a physical exchange of information. The combination of these two by the state is what essentially has become a burden for us. We have essentially become the right hand of the state acting as the mechanism for this digitized movement to move forward. It is especially tough right now because the economy of this country and in general is in a very bad spot right now. There are a lot of people that are half working or not working at all and buying for example a printer or a scanner is the last thing on their mind right now.

Do you see any problems in data sharing? For example, are you worried for legal reasons or for the protection of personal information, or are you worried for your competitors?

- I think that because all of this is new for a lot of people, we as in the majority, do not realize how important data and information sharing can be for someone else. You hear in the news that people can gather social and personal information about a person and use that information to access bank accounts. But we are still in a phase where we do not know how this can happen or how the collection of information can be used to do profilings on people. So all in all we are left with a big question mark as to what is safe to share and what is not safe.

Would you be willing to participate in a system where you can share private data without the specific details and without it being linked to you?

- Yes, I do not think that I would have a problem with that. Even now because everything is electronical nowadays we share a lot of stuff with other businesses. The only thing that we do not share and that they do not necessarily know is our stock.

Now that you mention your stock. Would you participate in exchanging information about what each bookstore has in their stock without you knowing the individual bookstore and without them knowing anything about the ownership of your bookstore but they would know what you have in stock?

- Yes, I think I understand what you mean, I think that this is personal information and it could be vital information. For example I could for some reason know that the other bookstore has a certain number of books and the next day that said bookstore might sell all these books. I would not know that the bookstore sold all these books. To give an example, I can know what a bookstore has in stock when I look at their bookshelves but I do not necessarily know what they might have in stock inside their storage room which is hidden from plain sight.

Would you like to know this detail?

- Personally, not so much, others that do not focus so much on their own store would maybe want to know details about your store. If you catch my meaning.. For me it is not appealing what the store next to me has in their storage facility.

If for example other bookstores sold this book by the dozens, would you personally want to know this?

- I think that this is a matter of supply and demand. We have a huge catalog of books and other items that we sell in our bookstore, some of these items are in high

demand and others are not. If there is a product that is in high demand you know for certain that you are going to reorder it from the supplier. If the supplier tells you that they are out of this product, it means that the product is in very high demand and all the bookstores are ordering it. What I mean to tell you is that we know which products are on high demand and which products are not. If another bookstore owner tells me that they sold three hundred stamp collecting books, I will not believe them of course… So yes that's all.

Interview 3

Now that I have briefly explained to you what this interview is about I would like to know what kind of an SME you own?

- A cafe bar on a small Greek island.

Is data sharing or information sharing a thing for you?

- It is, not on an IT level but mostly from mouth to mouth.

Are you participating in data sharing?

- Yes on an informal level.

If yes then how when with whom?

- It can be everything for the business, for how to run the business. It is done with other people that have similar but not competing businesses. Yes there is an exchange of ideas from products, to staff, to legal matters, so yes we share information about many subjects.

Can you specify, for example, wages, equipment, service?

- Equipment yes but we do not talk about wages we do not talk who is paying whom what. We definitely share other staff information like who to hire and who to not to. Definitely products or prices of products. Good places for you to get whatever products it is that you sell.

If you are doing it then what benefits are you getting?

- You get good ideas sometimes from somebody else who is in the same situation more or less and you keep a good connection to your colleagues, competitors. It is always good for business to have good relations with other businesses.

Do you think you could be getting more out of it, if so what, how, where. with whom.

- Definitely especially on a small island like this if everybody would share their experiences, their knowledge of products, prices, staff, wages, you could definitely help the situation for all businesses. But everyone has to join in and that is not always possible.

Is there any specific data that you would like to share or to receive?

- Right now laws around running a business are changing all the time. Laws about staff and the rules about the staff both wage wise and days off and all of this. It is changing fast and it is impossible to know all the details. So that kind of information if it would be more readily available then yes that would definitely be a help.

Is there anything that you worry about, do you see any problems with information sharing?

- Only if 90% are sharing fairly and 10% are just using everybody else's ideas. I mean if you share and somebody exploits the knowledge that you are all sharing in a way to get ahead of the others. But that is all I can see, otherwise I can only see benefits in sharing information about my kind of business or anybody's business I guess.

Would you be willing to participate in a system where you can share private data without the specific details and without it being linked to you?

- You mean like anonymous ideas or anonymous data sharing?

Yes sort off

- I mean yes why not if it is anonymous, anonymous ideas and it can help anybody, sure. I just do not see the reason for the anonymity but then again this is because this is a small place and everybody knows everybody. I guess in a larger environment anonymity would be preferred. But yes I would share anonymously or not.

Interview 4

Now that I have explained to you what this interview is about. What kind of an SME do you own or run?

- I run a cafeteria.

Does information sharing mean anything to you and if so do you partake in it?

- No, I do not partake in it.

Why is that?

- Because I do not want to share with anyone my information or secrets with other SMEs

Do you see any benefits to information sharing?

- No I do not think there are any.

Would there be any specific data or information that you would share?

- No, I do not think there are any specific details or information that I would share.

Are there any obstacles that come with information sharing, are there any problems with information sharing?

- If we are talking about trade secrets then yes. It is bad for me to share my ideas because if someone else steals them and puts them to use then essentially we will offer the same service which loses me customers. I would like to be the only one that has some service or product that is unique, so no I would not want to share.

Do you see any benefits to that? For example the creation of a better product or service or better partnerships?

- I think that there should be solidarity between SMES and more specifically cafes but everyone should do their own thing.

Would you be willing to participate in a system where you can share private data without the specific details and without it being linked to you?

- If it could convince me that I would have some profit out of it then yes.

Interview 5

Now that we have talked about what this interview is about I would like to know what kind of an SME you own?

- I advise and consult companies in their IT and software departments

Does information sharing mean anything to you and if so do you participate in it?

- If you mean between the companies that I work with then no because I can not share their information due to privacy and legal issues. Unless it is something with no identity, something generalized. For example I can not say that that specific company does this but I can say that this is the general direction of the market.

Okay and how do you share information?

- Usually I do this verbally because I visit my clients. We do not use so much email or other communication options because of the nature of this job.

Is it only SMEs that are of a different nature than yours or are you also doing business or sharing information with similar companies?

- Yes there are some times that we have shared information with competitors and we have even supported them. Especially here in our island the competition is small so everyone talks to everyone.

Ok so if you are participating in data sharing then what kind of benefits are you getting out of it?

- There is definitely better cooperation with the customer as they get better familiarized with our software. It also helps us to some degree although the information that we gather is useful to the customers and not so much for us. It is used to perform a better service as we can use their data to deliver a better experience to the customer.

Do you think you could get more out of information sharing?

- I think it has to do with the fact that we are a small community or market. We could for example sell some of our information but it is not simple to do so. It can happen to some extent but it does not make us profit. Maybe if we were on a bigger market then perhaps it would be a more profitable action.

Okay are there any specific data that you would like to share or receive?

- No I do not think so because my line of work has parameters that change all the time so I am not looking for something specific.

If you do participate in information sharing is there anything that worries you about that, is there a problem in information sharing?

- Nowadays with the implementation of GDPR, even though I do not think that it is implemented to a full extent, probably not. If you are not sure about something you just filter it out and you do not share it.

Would you be willing to participate in a system where you can share private data without the specific details and without it being linked to you?

- Yes I would.

Interview 6

Now that we have talked about the nature of this interview I would like to know what kind of an SME do you own?

- The enterprise that I own is a jewelry shop that deals also with the exchange of gold. This means that we also buy and sell gold and jewelry.

Does information sharing mean anything to you and if so do you participate in it?

- In the jewelry side of business, yes it does mean something as you need to know what the current prices are and in terms of buying gold you need to know on a daily basis as you need to know the price of buying and the price of selling gold. So it is something that I use every day.

If you do participate in information sharing how exactly do you do it, what are you exchanging, with whom and when?

- In terms of jewelry it usually happens through big expos and displays. There are big companies and medium sized companies that showcase their jewelry there. You can check their prices or learn what kind of jewelry designs will be good for the market for this season. You can learn the pricing difference between markets you know depending on the scale of the market. In terms of raw gold I usually find information online on sites that show the pricing every day or for the past hour day or week. So in terms of gold it is vital to have some information on a daily basis.

What kind of benefits do you get from information sharing?

- In terms of jewelry we get the benefit of being able to price them better. As it happens we are dealing with a smaller in size market so with this information we can be more

competitive. Because a lot of times most of the prices are fixed you find ways to be more competitive for example you can buy better quality items in order to justify your prices. You really need to know stuff because it is a difficult market to be in. In raw gold it is still competitive because it has to do with what you want to buy or sell. It really depends on the info that you have because you can look at how much you can lose now in order to profit later. You really need to know on an hourly basis the prices for gold. It really is a matter of survival.

Do you think you could be getting more out of it and if so what and how?

- Maybe if there was something more reliable and organized. In jewelry maybe not because everything has almost a fixed price. But in gold maybe if you knew more about the prices, to know if you can lose this week or buy at a lower price in order for the next month to sell higher. If there was some kind of service that would show you projections of how gold prices would develop.

What kind of information would you like to get and share with others?

- I would really like to know the purchases and sales of others, especially in SMEs what kind of prices they have I do not want to know the specifics for example to know exactly what the jewelry shop across the street sells but I want to know roughly how the market is moving in my town because I might be 15 or 20 percent down without knowing it thus lose value on every exchange.

Is there anything that worries you about information sharing or are there any problems that you face?

- I believe personal information is a problem. A good thing would be to exchange only numbers and not any identifiable information, especially in this market personal information should be protected in some way.

Would you be willing to participate in a system where you can share private data without the specific details and without it being linked to you?

- If it would help the growth and progress of my company, yes I would be open for a solution like this.

Interview 7

What kind of a SME do you own?

- I own a cinema which is located on an island.

Does information sharing mean anything to you and if so do you partake in it?

- Information sharing in our line of job has mainly two areas. One area focuses on the projection that helps us understand how many tickets we will sell based on the movie that we bring and the other is to find the right price for the acquisition of a movie.

How do you participate in information sharing for example how do you share it and with whom?

- Usually the acquisition or rental of a film is done through a company that distributes films. This distribution company either uses prognostics or through their knowledge they set a price in order for us to either buy the picture for a fixed price or by getting a cut from the tickets that we sell. Yes so basically this is the information that we mainly exchange. After each movie has run its course through our cinema we get in touch with partners throughout the whole country and we discuss whether or not each product or movie was a success or a failure.

What kind of benefits do you get from information sharing?

- Essentially we know by comparing with others if the movie that we show in our cinema has brought us some revenue or positive numbers. This helps us understand

the needs of our customers for example if they are into a particular set of movies. We can use that information to appeal to those needs and even further, become more competitive with our competition.

Do you think that you could be getting more out of this and if so, how, when and with whom?

- From information sharing I think that we could get a better base of information, for example to know if the price that we pay for a movie is competitive or not. Furthermore, to choose a better product in order to maximize profit and to be more competitive.

Are there any certain types of data that you would like to share?

- Yes for us it would be the prognostic type of data that would let you compare different themed movies in order to learn how they will fare in our market. It will help us understand what kind of profits we will earn and how much of a hit they will be. In turn we could make better decisions as to what kind of movies we should buy in order to make more profit.

If you do participate in information sharing, do you find any problems with it, is there anything that bothers you?

- I think that personal information and the specific details of every company or enterprise are a problem. Mainly because of competitiveness reasons.

Would you be willing to participate in a system where you can share private data without the specific details and without it being linked to you?

- Yes I think I would be willing to participate because it would be good for us to know what kind of profits we would make in comparison with other SMEs or cinemas in order for all of us to be more competitive with the distribution enterprises.

Interview 8

Translated from danish to english.

To begin with I would like to know how big of a business you are, because we focus on small to medium size businesses. So what is the approximate number of employees you have?

- We are about 80 employees

Ok, perfect. Then I would like to start with, do you share data with other companies or organizations?

- Yes we do. We share data with the company that takes care of our booking system.

Ok

- It is through our website where they convey our customers personal data when they make a reservation.

Ok. So you have an automated system set up for booking?

- Yes

Are you sharing data with others?

- No

Not with competitors?

- Not as a point of departure.

Ok. Now that you are sharing data with a third party. What kind of advantages do you see from sharing data?

- I mean you can say that, the cooperation with our booking system is not done to take advantage of data. It is done to take care of a service we can take care of ourselves, without too large an administrative load. So it is an automated process that has been established in pretty much the whole restaurant business.

Ok

- So in that way it is not to take advantage of data we do it, it is to help with our administrative workload. So that way they get access to some personal data, but they are also subject to the GDPR so they can't bring it forward.

Yes, of course. Would you be able to see any advantages in sharing data with competitors or business partners?

- Yes that could easily be, in relation to for example if we have a guest that does not show up to their reservation then we have a system that marks their data on our own. A guest has not shown up to their reservation, and so we have lost income at that table that night, because they didn't show up. If it could be a general thing that a person can become in bad standing, because they repeatedly didn't show up to their reservation, then it would clearly be in our interest as well as a competitors, that either, via a third party, we could share the information that they didn't show up, then they in the long run could be marked as bad guests, if you know what i mean.

Yes. So this is a type of data you could see advantages in sharing, but you don't do it.

- Yes, we take care of it ourselves. If there's a reservation and they don't show up and dont send an email, or calls, then we note them ourselves in our system that they have not shown up to their reservation, and there we can see a history that goes x number of months back according to GDPR. So it is't a history that can stretch back multiple years. I have worked with other booking systems that pre-GDPR could talk between the systems. So if they hadn't shown up at one restaurant then I could also see that because we worked under the same concern, and thereby had access to the same guest information.

Ok. But right now it only takes place internally?

- Yes.

So is it GDPR restrictions that means you don't share information with others right now?

- Yes because the company that takes care of our booking system does not develop it and is not allowed to develop it. Because it says very clearly with name, surname, email and phone number and other sensitive personal data on the reservation, and that means we are not allowed to share between each other. So as far as i know then it is not possible.

Are there other reasons besides the GDPR that you are not sharing data?

- I'm unable to say.

Ok. So the last thing is, one of the things we are looking at right now is different methods of sharing data. For Example that it can be done anonymously, and in a way so others can't see specific details but can see general data. So you would be able to see data from multiple restaurants, but not more specific data but more averages and such. Would that be something you would be interested in?

- Yes and no. It depends on the size of the restaurant I think. In a restaurant like the one I'm working at right now, where I'm the one taking care of all the bookings. I would say it is very limited the amount we go in and screen every single reservation, or get into every single one. We go through it once a day to see if there are any VIPs, but we know that beforehand, because we run an internal log of, where we can mark it ourselves. And what is important for one restaurant is not necessarily important for all restaurants, so if I get information from another restaurant that doesn't have the same work process as us, I can see some difficulty in. So over sharing can also be a problem, and unnecessary. So not 100% yes or 100% no.

Interview 9
Translated from Danish to English

So this is about datas sharing between different businesses, we are focusing on medium to small businesses, so how many employees do you have in your business?

- You will have to specify

Is it under 250?

- Yes

Ok that's good enough. I want to start by establishing whether or not you are sharing any data or information with other businesses.

- No we don't actually.

Ok, you don't share with competitors, partners or other organizations?

- No

Ok that's alright.

- By data you mean if we have some form of digital system?

It can be digital or personal, all kinds of.

- Ok then we. I'm part of two restaurateur coffee clubs and in a professional restaurateur association, and there we share data, but we don't have anything automatic.

Ok. We are investigating if data sharing takes place and how it happens.

- Ok

What type of data are you sharing when you share?

- Well then it becomes very personal. There's probably not a lot who share. Is it within all types of businesses or is it specifically the restaurant business?

Right now we are looking specifically at the restaurant business.

- Ok then there's probably not many who are sharing anything. There's not that many restaurateurs so there's of course some that you know better than others, and there you just share some things such as the salary percentage, and profit margin, and such things.

Ok.

- And things like the gross profit. And turnover, but that's not so interesting.

Ok, we are just trying to figure out what is going on.

- Yeah and stuff like "what do you pay a new employee without experience?". That is also a type of data. Just back and forth.

Ok so you are saying that you don't have a specific system set up. Are there any reasons why you don't have that?

- Yes, I think most people just want to keep it secret. I can't tell you why, but that's the reason why you don't share it.

So it's private?

- Yes, I think that's the way that people think.

Do you think you could see any benefits if you could share that type of data?

- Yeah there would probably be some. Fyens Stiftstidende has just begun within the last couple of years to think it is fun to write about restaurant businesses, which they dont do with any other businesses, but about their accounts. And there we are some that have sole entrepreneurship, and that they of course can't get access to. And we are pretty happy about that, and those that don't have sole entrepreneurship are not happy about it, no matter if it is a good or bad finances. Because you can't win anything from it by getting it in the press. So if you are going to be sharing anything

then it will have to be internally secret. If you know what I mean. If it goes too well, then people think it is probably because it is too expensive, and if it is going poorly then people stop coming. So you can't really win anything by having your finances in the press.

Ok

- And I think that lies as a fear and is why you don't share that much.

Ok. We are specifically looking at these reasons why data is shared and why it is not shared. And we are looking at ways to share these, private or secret data, without it being traceable back to the specific business.

- Ok. Then you have got something there. Then you can. If it is anonymized to some level then it begins to be interesting.

Ok s then that you would like to

- If that was the question then yes.

Ok. Because right now we are looking at something where you can share secret data without it being traceable back, and you can see it from others, but you can't see specific data. You can see more like averages and general data but you can't see specific data from specific businesses. Would that then still be

- Yeah, then you just have to be sure that it is categorized correctly, before it has value, because there is a giant difference between McDonalds and Noma.

Yes

- But otherwise yes.

Interview 10

Translated from danish to english

Ok. We are interested in small to medium sized businesses. So to begin with I just want to ask, about how many employees you have?

- I would say about 25

Ok, perfect. So to start with I would like to establish whether you are sharing data. Whether you are sharing data with other businesses or organizations.

- We do not.

Ok so you don't share information with organizations, competitors or business partners?

- No, it is two brothers that have it.

Ok. Are there any reasons why you are not sharing data?

- Not as far as i know

Do you think you would be able to see any advantages to sharing data with business partners or competitors?

- I guess I actually do think there could. Yeah on some business, about what the trends are on the market.

Yes

- Then it could be a good idea. And then again not. Our concept is too strong in the local area that we don't see any reason to.

Ok.

- But we do go around to other restaurants, and let ourselves be inspired by their menus and interior design, and so on. So that we can keep up.

So you don't share data directly but you

- It's more like spying.

Ok. So at the moment we are looking at different techniques for sharing data, and specifically techniques that maintain privacy. So we are looking at a technique that allows you to share your data, and you can see others' data, but you can't see specific data, you can only see general data such as averages, and totals. Would that be something you would be interested in?

- I don't think so.

You said that you would be interested in trends and such.

- yes , it could of course be interesting, but there we just have, we of course let ourselves be inspired by social media mostly. And then as said, we go around and sniff around places that are popular.

Ok so the techniques you have already are good enough, you don't need additional

- Yeah we also don't spend any money on advertisements.

Ok

- So both our facebook and instagram pages are pretty quiet.

Ok

- We are going more for the weather, word of mouth. And then we are located in an area where there's a lot of apartments around, we are located in the middle of a housing block for eksempel. So we are nice and easy to get to for the locals, and have actually that way full seating every evening in the week

Interview 11

Ok so, we are doing a study on data sharing among small to medium sized businesses. So first of all, I would like to ask roughly how many employees do you have?

- I would say, including all the yoga teachers and the cafe staff, maybe about 25.

Ok excellent. So we want to establish whether small to medium businesses are sharing data with competitors or organizations or supply chain members. So is your business sharing data?

- So i was thinking about it when you emailed me and i was like, i don't really think we purposely do. We of course like to have third party software that we use for yoga classes and we have the POS system for the till that has an insight into everything that we kind of sell, so they have that. And we also have some, like CRM systems in place to get in touch with our customers. But we don't really specifically send it anywhere, if you know what i mean.

So you have some sort of sharing, but not explicitly going out of your way to share data.

- no

Ok. You talked a bit about the kinds of data you share and with whom.

- Yeah. Do you want me to tell you more?

Sure yeah that would be great.

- We have got. Because we are a company that has both a yoga studio and a cafe, the systems are kind of not connected. We use this third party app and system called Mindbody, it's an American thing that we like to use for booking and memberships and all yoga related stuff. And then attached to that we also have something called Loyalsnap. Which is kind of like an app that scrapes all the data from Mindbody and allows us to stay in touch with our customers. So those two kind of correlate with each other and send each other like data I guess. And that is stuff like what kind of memberships people have, like peoples phone numbers, emails and like how many times they have attended and bla bla bla. So we have an overview of like, whose membership is ending, we call to like bring you in bla bla bla. And then we also have

like the sales system from the cash register in the cafe, which is automatically shared with the people that run the system that we have in it. And they can log into it and change stuff for us if we want to. And then we also have tracking data like google knows and facebook knows because we do like paid ads and stuff, so they have the data from that, but we don't send it to them, they just collect it so they have it. Yeah that's it.

Ok great. So what kind of benefits is it you are seeing from the kind of data sharing you are doing?

- In our case. I think the fact that we have those two yoga systems that are connected, that we don't have to input anything ourselves that they kind of share that data between each other. So we don't have to manually keep track of anything, so it is mostly for efficiency and accuracy i would say. To reduce human error because it is all systems that scrape data themselves. And with the cash register I would say I don't know if I see any actual benefit. It's maybe just security because they know how to set everything up properly. They can advise us on that, and apart from that I don't think there's any specific benefits.

Ok. Do you think you could find benefit in sharing other kinds of data?

- Maybe if we were into some kind of consultancy or someone who would give us advice on how we could improve our yoga system, or our sales in the cafe. That's really the only way that I could see it benefitting our business model realy.

Ok. What about supply chain partners or competitors?

- I guess like with the groceries for the cafe, we do have a couple of suppliers that we work with and we of course kind of tell them, oh this supplier charges us like this much can you match that price? I don't know if that classifies as data sharing. But we don't really. We kind of want to stalk our competitors rather than give them our data in that way. We want to know what they are doing, but not have them know we know what they are doing. So it's more of, like, a secretive non-sharing data kind of environment.

So you are interested in receiving data from others but you want to sort of protect your own.

- Yes, of course. But I don't know if that balance is possible, really. But we do kind of try to fish out what others are doing, but it is mostly like. I wish I could see the statistics of our competitors in the area of, like how many memberships they have sold in the last month of course, to see to compare, but I don't think I will ever have access to that. But yeah

Ok so the things that are stopping you from trying to share data is you want to protect your own.

- Yeah. Yeah probably to protect and to make sures that. Because we have had a couple of instances where some of our competitors around us have kind of copied some of our campaigns, but like made them more attractive for the customer. So, like making them, like, 50 kr cheaper or something. So we have had those cases and it is really annoying because it is a really aggressive way of marketing and campaigning, but yeah we have had those cases. So we don't really like it when people know our stuff.

Ok. So one of the things we are looking at specifically is different techniques for sharing data, but doing so in a way where you can protect your privacy and confidentiality of your information.

- Ok

So we are looking at a system where you can share data but the data can not be traced back to you. So you can share data with others but you can't see specific details, you can see sort of trends and averages and totals, but you can't see specific data about anyone. Would you be interested in such a system?

- That would probably be useful. I think I mean, to have an overview in general of the data, like, some statistics of the industry and things like that, but not being able to trace it to specific companies, could be quite useful, in strategizing and planning future campaigns and stuff like that. So that would definitely be interesting.

And would you be willing to share your data with such a system?

- If people didn't know it was us then yeah.

Interview 12

Translated from Danish to english

Ok so, We are doing a studio on data sharing between small to medium sized businesses.

- Yeah

So I want to start by asking ca. How many employees do you have?

- In total here or in the group I'm in?

Total in the group

- I don't actually know. I would guess about 30 to 40.

Ok that's perfect. Do you know if you are sharing data in any form?

- We don't do that.

You don't share with business partners or organizations?

- Yes in the organization we do of course. All the data from the different places is collected in one place.

But it is internally?

- Yes it is internally.

Do you think you could see any advantage in sharing externally?

- No, not immediately.

Not by sharing with business partners?

- I guess we could. There are some of our suppliers, we have Carlsberg for example as a supplier, and of course they would probably be able to gain something by seeing our, but they can already see that, in correlation to the amount we are buying, about how our consumption is varying.

Would you be able to gain something from seeing data from others?

- Yeah of course we would. Because when we are starting something new, a new restaurant or something. Then we would be able to get, if we had data about how much is sold and how busy it is in different periods. Then we would be able to plan much better and more efficiently according to that.

Ok. So are there any specific reasons why you are not sharing data?

- It is simply because we don't have the resources for it. Resources are not being set aside to take care of that side of data, whatever you call it.

Is it because of the size of

- Yes. It is because of the size and because it is not something that is thought about so much in the organization.

Ok. We are working specifically with where you can share data with other businesses, where you cant see specific data but, so if you are sharing data, it would not be traceable back to

you. But you would be able to see averages and totals and such. Would that be something that you would be interested in?

- Yes. it definitely would. Again because it would also be able to benchmark us relative to how others are doing in the same group in this period. And then we can see if we are above or under, and then adjust accordingly, or see what we could then do.

Ok. But the problem is the size and you don't have the manpower

- Yes but, the bottom line is that we don't really have the manpower for it. It might be what I end up sitting with. But the bottom line is there's not the manpower and there's also that often the restaurant business is relatively old fashioned. The whole business with new stuff can take some time before it catches on.

Interview 13

As we are talking about SMEs I would like to start things off by asking how many employees you have?

- We have eight to ten employees. It depends on the season.

Okay, are you partaking in any data sharing, either with competitors or some of your partners?

- We are sharing data with some of our competitors.

What about some of your partners?

- As far as I can tell no we do not. Although one thing that comes to mind is that when the customers book their table it is done through an enterprise that handles the booking reservations. This means that they have their information in their database and so do we.

Okay what about with some organization branches?

- We are part of the Danish restaurant and cafe branch but we do not share customer data between us.

What about internal information or data?

- What do you mean?

Like maybe salaries?

- Yes we do share that information with the salary bureau, we keep data on our employees.

So you share customer data with the booking organization and salary data with the salary bureau?

- Yes we do and for example when we do share customer data we can not do so much about it because they have to go through the booking system.

Do you see any benefits when you share information?

- I mean it is not something that we can avoid, otherwise everything would need to be manually arranged. We do not do it for a specific reason, we just do it this way.

Do you think that you would benefit more if you shared with more organizations, branches or competitors?

- Maybe if we could arrange some partnerships. For example we have made in the past some arrangements and events with some wine distributors. It could be beneficial if we could send email to the customers that would be interested in that sort of an event. Alas it has not been that relevant in the last couple of years.

Is there any reason that you are not sharing data?

- No, I do not think that there is a reason behind it. I guess we are busy enough with just making food and serving it. We are just a restaurant and not so much a software company.

Right now we are working in different ways of sharing data with other restaurants or other enterprises.

- Is customer data, because we can not share this type of data. For example I can not share a customer's email address with a competitor.

The idea is that you can make the information anonymous this way it can not be traced back to anyone.

- It sounds like it is a bit illegal.

The idea is that it will be done in a way that it can follow the law in terms of privacy. In technical terms, you would not be able to identify any specific people, organizations or enterprises. You would however be able to share information, it is a new technique that we are looking into.

- When the customers insert their data into our system it is only to be used for the function intended for example only for the table that they have booked. I think it sounds a bit unethical. I am not sure, are you researching this or are you developing some software?

Right now we are looking at some techniques that will enhance privacy when we are sharing information but before that we would like to find out if there is a need to do this.

- The short answer to this is that we have too many customers that we have to take care of. Either they book a table or just drop in. We do not need that kind of a system. Maybe if we were to arrange events then it would be really nice to use such a system to attract more customers. I have to say that we are also advertising through facebook and when we upload something the data that is exchanged there is on a whole other level but again it is not in our hands. We would not actively participate in customer data sharing. I do not think that it would be fair if for example I look at a restaurant and after a while be bombarded with advertisements.

What about other types of information, for example information regarding products? For example wages or information on product prices.

- For example prices from our distributor then we could take all the distributors and compare prices?

Yes, for example, what about that kind of information?

- Well in what reality would you have the time to sit and go through all that information we are just a restaurant.

We can see that other bigger enterprises use and interact with that type of information.

- Bigger enterprises are a completely different story. I have a few hours that I use to look at bills and salaries. I do not want to use more applications to send and receive data back and forth. It is not realistic that the size of our organization can use so much effort into that.

Okay that is great to hear, it is not something that you have the resources to pull off.

- I do not think so, we decline arrangements and we get new distributors that try to sell us new products and maybe they have some good prices but we can not spend time on exchanging all the time with our product distributors. It is a lot of work to exchange distributors. It is not worth it to do all that if it is to just save a couple of crowns here or there.

Okay that was everything I had

- Okay well best of luck with your project

Thank you, bye.

- Bye.