Master Thesis at the Institute for Computer Science at Freie Universität Berlin

Research Group Human-Centered Computing (HCC)

A data visualization tool for finding optimal privacy-utility trade-off in data analysis

Zheng Yao

Student Number: 5262478

zheng.yao@fu-berlin.de

Supervision: Dr. Daniel Franzen

First Assessment: Prof. Dr. C. Müller-Birn

Second Assessment: Prof. Dr. Volker Roth

Berlin, February 21, 2025

Eidesstattliche Erklärung

Ich versichere hiermit an Eides Statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben, Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Berlin, den February 21, 2025

Zheng Yao

Abstract

Data dissemination is prevalent and crucial in contemporary times, yet concerns about privacy breaches compel data controllers to consider methods that safeguard personal information when releasing data. Data anonymization methods offer privacy protection but come at the cost of sacrificing data utility. Balancing privacy and utility has been a concern for data controllers. Traditional approaches heuristically determine anonymization methods based on past experiences to achieve anonymized data that meets privacy requirements. However, these approaches predominantly consider privacy and lack substantial research on utility metrics. This study proposes a novel approach to balancing privacy and utility in data anonymization, integrating both aspects within the same framework and visualizing the trade-offs of privacy strategies. Our study introduces 'VisTool', a prototype designed to quantify privacy and utility losses during the data anonymization process, offering intuitive visualizations to aid users in making informed anonymization decisions. Through the integration of various visualization methods and user interface components, we successfully amalgamate the trade-off results between privacy and utility. This integration enables users to gain a clear understanding of different aspects of anonymization strategies, providing them with intuitive and transparent trade-off outcomes. In the current research landscape, there is a limited focus on the visualization-supported trade-off between privacy and utility in data anonymization. Our research aims to address this gap by delivering a comprehensive and user-friendly anonymization decision tool for data controllers. Through evaluations, the paper demonstrates the functionality and usability of the prototype, providing evidence of its effectiveness in achieving the desired balance between privacy and utility.

Die Datenverbreitung ist heutzutage weit verbreitet und von entscheidender Bedeutung. Bedenken hinsichtlich Datenschutzverletzungen zwingen Datenverwalter jedoch dazu, Methoden in Betracht zu ziehen, die personenbezogene Informationen beim Freigeben von Daten schützen. Methoden zur Datenanonymisierung bieten Datenschutz, gehen jedoch auf Kosten der Datenverwendbarkeit. Das Abwägen von Datenschutz und Datenverwendbarkeit ist eine Herausforderung für Datenverwalter. Traditionelle Ansätze bestimmen heuristisch Anonymisierungsmethoden basierend auf vergangenen Erfahrungen, um anonymisierte Daten zu erhalten, die Datenschutzanforderungen erfüllen. Diese Ansätze berücksichtigen jedoch hauptsächlich den Datenschutz und weisen kaum Forschung zu Metriken für die Datenverwendbarkeit auf. Diese Studie schlägt einen neuartigen Ansatz vor, um Datenschutz und Datenverwendbarkeit bei der Datenanonymisierung auszubalancieren, indem beide Aspekte in denselben Rahmen integriert und die Trade-offs von Datenschutzstrategien visualisiert werden. Unsere Studie stellt 'VisTool' vor, einen Prototypen, der entwickelt wurde, um Datenschutz- und Datenverwendbarkeitsverluste während des Datenanonymisierungsprozesses zu quantifizieren und den Benutzern intuitive Visualisierungen zur Verfügung zu stellen, um informierte Anonymisierungsentscheidungen zu treffen. Durch die Integration verschiedener Visualisierungsmethoden und Benutzeroberflächenelemente vereinen wir erfolgreich die Trade-off-Ergebnisse zwischen Datenschutz und Datenverwendbarkeit. Diese Integration ermöglicht es Benutzern, ein klares Verständnis verschiedener Aspekte von Anonymisierungsstrategien zu gewinnen und ihnen intuitive und transparente Trade-off-Ergebnisse zu bieten. In der aktuellen Forschungslandschaft liegt der Fokus begrenzt auf der durch Visualisierungen unterstützten Abwägung zwischen Datenschutz und Datenverwendbarkeit bei der Datenanonymisierung. Unsere Forschung zielt darauf ab, diese Lücke zu schließen, indem sie ein umfassendes und benutzerfreundliches Anonymisierungsentscheidungstool für Datenverwalter bereitstellt. Durch Bewertungen zeigt das Papier die Funktionalität und Benutzerfreundlichkeit des Prototyps und liefert Beweise für seine Wirksamkeit bei der Erreichung des gewünschten Gleichgewichts zwischen Datenschutz und Datenverwendbarkeit.

Contents

List of Figures

List of Tables

1 Introduction

Every day in our lives, enormous amounts of data are created. We have completely submerged ourselves in the information society. Most human acts leave digital traces, which are collected and stored by someone. Social media, the Internet of Things, bank transactions, and retail purchases are just a few examples of ways to collect data. A huge data collection offers several advantages, including improved economic opportunities, better and more rigorous research, and in general, brighter prospects for increasing the human race's well-being.

However, gathering, sharing, and publishing personally identifiable information (PII) has a worrisome side effect in that it invades the privacy of the persons to whom the PII relates; a well-known example is the teenaged pregnancy guess published in [Duh12]. Data protection law, exemplified by the EU General Data Protection Regulation [Wik22], aims to safeguard individuals by limiting the accumulation of PII. Anonymizing PII, or converting it into data that is not individually identifiable but has significant analytical utility, is a method of enabling data analysis, sharing, and even publishing without violating data protection laws.

1.1 Context and Motivation

In the domain of anonymization for privacy-preserving data publishing, we define the following involved stakeholders. The usual setting in anonymization is for a **data controller** (the entity that manages and releases the data and often owns them) to hold the original data with the original responses by the **data subjects** (Individuals whose data are represented in databases or are collected by applications.) and modify them to reduce the disclosure risk. Then the data controller publishes the anonymized data or shares them with **data consumers**, typically data analysts or researchers, who expect the anonymized data to be still useful. It may occur that some of the data consumers behave as **data attackers** and try to perform disclosure attacks on the anonymized data and breach privacy. Disclosure can be of two types:

- Identity disclosure, whereby the data attacker determines the data subject to whom an anonymized data item corresponds;
- Attribute disclosure, in which the anonymized data help the data attacker estimate the value of a confidential attribute for a certain subject.

The specific definition will be given in section 2.2.1.

Data at the individual level, such that each record corresponds to one data subject (person, enterprise, etc.), are known as microdata. From microdata,

other formats can be derived, such as tables and online queryable databases. Here, we will focus on microdata.

The traditional approach to anonymization is: The data controller runs an anonymization method (e.g. k anonymity[LWFP06]) with a heuristic parameter choice. After that, the data controller measures the risk of disclosure, which can be done empirically by attempting record linkage between the original and the anonymized data sets[TDF03]. If the data controller thinks the remaining risk is too high, he/she re-runs the anonymization method with more confidentiality-stringent parameters and probably more utility sacrifice. It is difficult for the data controller to assess the utility and privacy trade-off analytically [DP91], hence empirical comparisons are frequently used[DFT01]. That gives the motivation for our work, designing a system for helping data controllers balance the privacy-utility trade-off analytically.

1.2 Researching Goal

The primary purpose of this work is to create a tool that may be used for helping data controllers find the right utility-privacy trade-off to analyzing and publishing anonymized microdata. In order to do this, we consider the following research question:

Main RQ: *How can a tool support data controllers to find a right balance in the privacy-utility trade-off when anonymizing microdata?*

To answer the main research question, two subsequent questions should be considered. First, we should clarify the aim of the tool is to help data controllers find the optimal balance of utility-privacy trade-off. However, there is still no standard currently to define what is an "optimal" trade-off. In general, an "optimal" balance for anonymizing microdata is a solution that maintains maximum utility (the information in the microdata) while minimizing privacy disclosure. Therefore, our first research question is how to compare utility with privacy. Second, we think that user interface (UI) components and data visualizations will have a positive effect on finding the optimal balance. However, which UI component is helpful in such a tool must be discussed in detail: which interface elements can improve the usability of a tool that helps data controllers find the optimal utility-privacy trade-off?

RQ1: *How should data controllers consider privacy and utility in an integrated framework for privacy-preserving data publishing?* We currently lack a framework for thinking about the privacy-utility trade-off in data publishing. The majority of existing work on privacy-preserving data publishing uses the following heuristic approach. First, one chooses a specific privacy requirement, such as k-anonymity, l-diversity, and t-closeness based on intuitions of what privacy means. Second, one considers the following problem: after fixing a parameter iterative for the privacy requirement (e.g., tune k from 2 to 10 in k-anonymity), how to evaluate the utility of the generated dataset so that we can maximize a particular utility measure, which can be the number of equiv-

alence class, or the discernibility metric[LDR06]. The preceding approach is constrained in its consideration of the trade-off between utility and privacy. We have two main issues to solve: First, how to choose among the different privacy methods. Second, how to choose a particular parameter for the particular requirement? In this approach, these concerns are examined only from the perspective of privacy, with no regard for utility. However, this is insufficient since oftentimes there is no established privacy need in place.

In this thesis, we design a new approach to balancing the utility-privacy trade-off. The innovation of this approach against the traditional one is shown as follows:

Inno 1. **Automatic privacy criteria control (Auto PCC)**. We apply the anonymization method to the original data iteratively. The anonymization method used here is a combination of k-anonymity, l-diversity, and t-closeness. To simplify the steps of adjusting the parameters of the anonymization method, we used a privacy criteria control (PCC) proposed by Chou[CWM16], which derives a unified parameter p, that integrates the three privacy parameters (k,l, and t) into one single intuitive value. The detailed method of this PCC will be introduced in section 3.1.2. Since the parameter p has a value domain from 0.0 to 1.0, we can let the tool iterate the anonymization process automatically with the value p from 0 to 1, which means no privacy control to the strictest privacy control. We set the number of loops and after this process, a mass of samples with different privacy requirements will be derived. We assume that this intensive parameter selection will give the users more possible solutions than the traditional approach.

Inno 2. **Consideration of privacy and utility at the same time.** The utility and privacy of each sample will be quantified. The measurement will be calculated as the information loss of utility and privacy with different measurement methods (introduced in section 3.2). Then we visualize the information loss of both utility and privacy using a **scatter plot** at the same time, where the y-axis and x-axis represent utility loss and privacy loss respectively. Data Controllers may select one of the optimal points as the output for data publishing.

To address RQ1, we utilize a case study to show the result of the new approach. First, we will apply several data sets with different features to our new approach (detail is shown in section Section 4.1). We will observe the distribution of the points on the scatter plot to see if we can find the optimal trade-off. The degree of dispersion of the samples may also show if our Inno 1. is helpful by sampling or not. Since we have not experimented, We proposed some possible outcomes of the utility-privacy trade-off as a reference. Figure 1.1 shows the distributions of 100 imitated samples which represent the utility-privacy points. These points are generated randomly which fit the

power distribution under different parameters a. The real result of the experiment may fit other distributions. We just take Figure 1.1 as a possible instance.

In Figure 1.1, we can see when the value a is small (e.g. a=0.01 or 0.1), these 100 points are clustered and overlap each other, which means the iterations in Inno 1. are redundant. Moreover, the points are clustered near the origin, which means it would make almost no change to the result no matter which point we select. In other words, almost all of these points can get a "right" solution. When value a is large, for instance, a=1, the samples are evenly distributed. In this case, we know that almost every iteration of the parameter p does make a different trade-off result. When value a is even larger (a=5 in the example), the points are clustered far from the origin, which means the outliers that are near the origin are the optimal solutions. In short, a higher value a indicates that our new approach is more significant than the old one, and a lower value means the new approach will not help too much with anonymized data publishing.

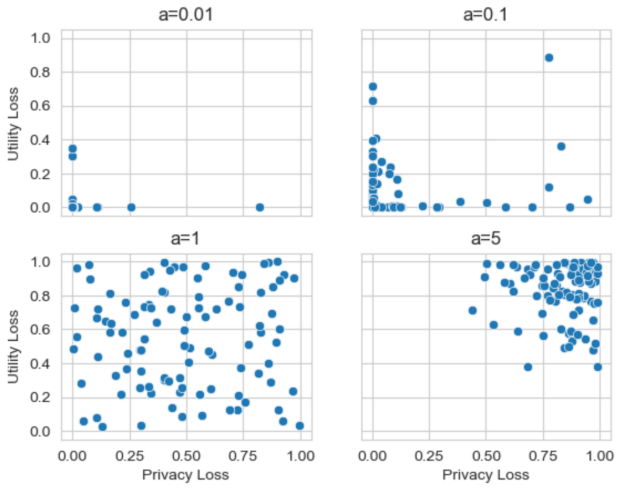


Figure 1.1: Hypothetical Outcome Plots of Utility-Privacy Trade-off

In conclusion, the proposed innovations, Auto PCC and the scatter plot of privacy and utility aim to enhance the effectiveness of addressing RQ1. The Auto PCC provides users with a diverse range of potential solutions, offering flexibility in meeting various privacy requirements. The scatter plot of privacy and utility losses serves as a decision-making aid for Data Controllers.

The hypothesized outcomes of the case study scenarios, illustrated in Figure 1.1, suggest that a higher parameter value (e.g. a=5) indicates the increased significance of our approach. In such cases, the clustering of points away from the origin signifies more meaningful trade-offs, emphasizing the tool's potential impact on decision-making. Lower values of a suggest that the new approach may have limited influence on anonymized data publishing.

In summary, these innovations are expected to empower Data Controllers

with a more user-friendly and comprehensive tool for navigating the privacy-utility trade-off. Auto PCC and the scatter plot are anticipated to provide valuable insights, aiding users in making informed decisions aligned with their specific needs and preferences in data publishing.

RQ2: *which interface elements can improve the usability of such a tool that helps data controllers find the optimal utility-privacy trade-off?* In RQ1, we proved through a case study that our proposed Auto PCC and scatter plot have a positive effect on finding the correct trade-off between utility and privacy. But at the application level, we need to design a tool that satisfies the usability of data controllers. We initially proposed the following UI components and utilized comparative experiments to find which components are suitable for such a tool. That is, which components have a positive effect on improving the usability of such a system?

1. **Spreadsheet**. In the traditional heuristic approach to finding the optimal trade-offs between the utility and privacy of anonymized data, observing the anonymized data is a trivial way to help data controllers make their anonymization decisions. We use the spreadsheet as a baseline in the comparative experiment.

2. **Trade-offs scatter plot**, which is the main component we want to prove the usability of this tool. It shows all trade-offs under different p values (conducted by PCC).

3. **Numierical indicators of Privacy and Utility loss**, which present the privacy and utility loss of a single trade-off sample numerically. The method of utility and privacy measurement will be discussed in Section 3.2), we think that this component will support the users in making the anonymization decision.

4. **Interactions**. It is mainly reflected in business logic regarding the choice of anonymization strategy, including anonymization method and parameter selection. We have three methods to do this: 1. **Trivial method**, which indicates the traditional anonymization strategy selection. The users have to manually select anonymization methods and their respective parameters. 2. **PCC method**. Compared with the trivial method, the PCC method simplifies the selection of parameters but still requires manual adjustment by the user. 3. **Auto PCC method**. The PCC method is continued, but users no longer need to manually adjust parameter selection, all processes will be performed automatically.

These components are only the main elements used in the comparative experiment. Detailed components will be added along with the development of the tool.

The comparative experiment will be conducted as a usability test with data controllers. We measure the usability of the UI components in three perspectives, namely **effectiveness, efficiency, and satisfaction**. These three

aspects will be measured by questionnaires Section 2.3 filled by the test users. Due to the limitation of finding plenty of participants for the usability test, we utilized the "think-aloud" method among dozens of participants to conduct a qualitative analysis. Although dozens of participants can not provide enough data for quantitative analysis, the results of the questionnaire can still demonstrate a trend, which will provide a direction for future work.

We hypothesize that our proposed methods, such as Auto PCC and scatter plot, may offer higher usability compared to the traditional approach of using a spreadsheet for direct observation. Through the comparative experiment, we aim to gather evidence supporting this hypothesis and guide the iterative design of the tool, ensuring an integration of UI components and interaction methods. The results of the experiment will serve as a foundation for refining the tool's interface, ultimately delivering a user-friendly solution for data controllers navigating the privacy-utility trade-off in anonymized data publishing.

1.3 Contributions

The primary objective of this study is to address the challenge of balancing privacy and utility when selecting microdata anonymization strategies in the context of privacy protection. To achieve this goal, we propose a visualization-supported tool designed to provide data controllers with intuitive measurements of both privacy and utility. The innovation of this research lies in two key aspects: firstly, the automated extension of Chou's Privacy Criteria Control (PCC) method[CWM16], offering a more comprehensive selection of anonymization parameters to meet a wide range of privacy requirements. Secondly, the integration of measurements for both privacy and utility through the adoption of various UI components, providing data controllers with a more intuitive approach to anonymization strategy selection compared to traditional heuristic methods.

Through the implementation of a case study, we have verified the functionality of the proposed tool and assessed its performance on datasets with different characteristics. In usability testing, comparative experiments with traditional methods have confirmed the enhanced usability of our proposed tool. In the academic domain, this research contributes supplemental insights into the parameter selection of Chou's PCC method[CWM16], validating its effectiveness. In practical applications, the tool is anticipated to assist data controllers in customizing privacy strategies across different domains, thereby improving the efficiency of data anonymization decision-making in real-world scenarios.

1.4 Structure of the Thesis

In this section, we provide an overview of the main concepts that drive our research. Furthermore, we outline the structure of the paper to guide readers through the subsequent sections. As Figure 1.2 shows, this thesis consists

of six chapters. In Chapter 1, we introduced the motivation and context of the research and articulated the central research questions. We outlined the specific objectives that aim to achieve clarity and depth in addressing these questions. In Chapter 2, we provided the basic definitions in the context of data anonymization and presented a brief review of existing problems involved in our research. Furthermore, after introducing related works, we highlighted gaps and challenges that our study seeks to address. A detailed exposition of our research methodology is also presented in this section. We described the research design, data collection methods, and analytical techniques employed to derive meaningful insights. In Chapter 3, we introduced the implementations of algorithms and methods we utilized for data anonymization and parameter selection. We described the quantification methods of utility, privacy, and the utility-privacy trade-off, which is the key point in measuring our research result. Additionally, we introduced our low-fidelity and high-fidelity prototypes. Chapter 4 is the evaluation of our research, Which ensures the reliability and validity of our findings in the case study and usability test. A discussion with data experts will also be introduced in this chapter. Chapter 5 is the discussion, concerning the result related to research questions of this study, providing context and deeper insights into the findings. In Chapter 6, our paper concludes with a summary of our key findings, a reiteration of our contributions to the field, and reflections on potential avenues for future research. The conclusion encapsulates the essence of our study and underscores its broader impact on the subject matter.

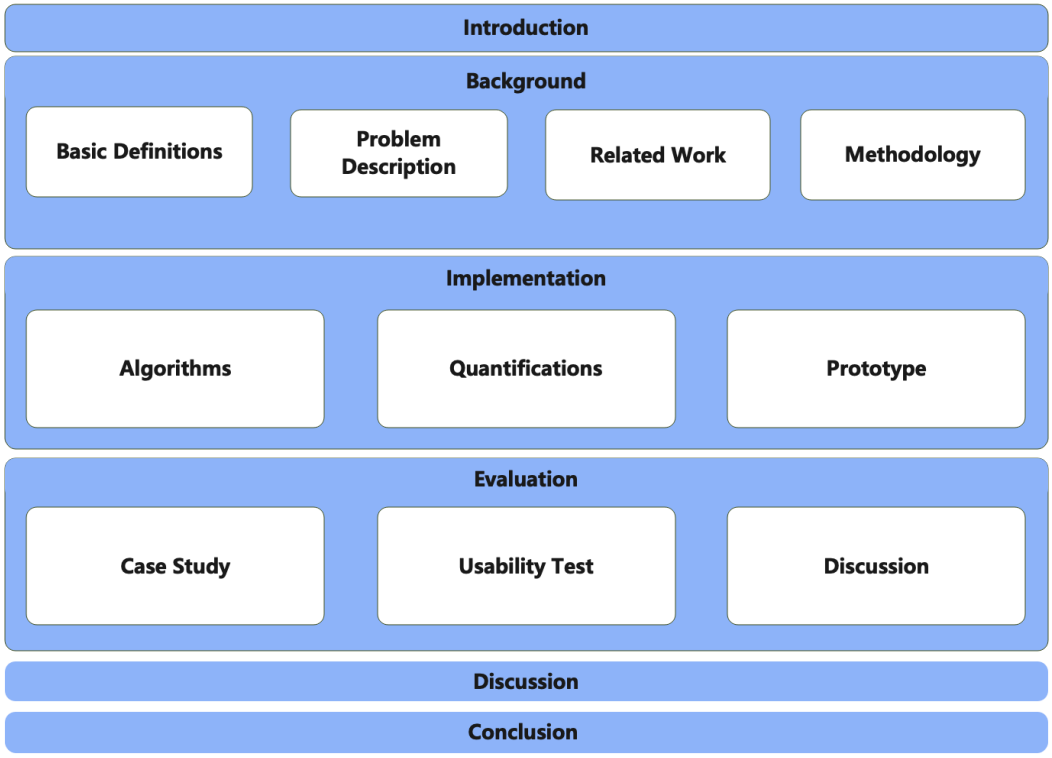


Figure 1.2: Structure of the Thesis

2 Background

The broad issues in the research are discussed in this chapter, with a focus on the terminology used in the domains of data anonymization as well as some techniques regarding the visualization of anonymized data.

2.1 Basic Definitions

In a dataset. All attributes are divided into 4 categories depending on their properties:

(1) **Identifier**[BA05] Attributes that identify an individual, such as ID number, name, etc., which must be removed from the published data set;

(2) **Quasi-identifier(QI)**[LWFP06] Attributes that can be combined to uniquely identify individual information, such as Age, Gender, Zipcode, etc., which need to be anonymized;

(3) **Sensitive attribute(SA)**[LWFP06] Attributes containing individual privacy information, such as Disease, home address, etc., will not disclose privacy when it cannot be associated with an individual;

(4) Other properties do not belong to the above 3 types of attributes, such properties do not need to be specially handled.

2.1.1 K-Anonymity

Gender	Age	Pcode	Problem
male	middle	4350	stress
male	middle	4350	obesity
male	young	**4351**	stress
female	young	**4352**	obesity
female	old	4353	stress
female	old	4353	obesity

Gender	Age	Pcode	Problem
male	middle	4350	stress
male	middle	4350	obesity
*	young	**435***	stress
*	young	**435***	obesity
female	old	4353	stress
female	old	4353	obesity

Gender	Age	Pcode	Problem
*	middle	435*	stress
*	middle	435*	obesity
*	young	435*	stress
*	young	435*	obesity
*	old	435*	stress
*	old	435*	obesity

Figure 2.1: (a)left: raw table. (b)middle: 2-anonymous view by local recording. (c)right: 2-anonymous view by global recording. [LWFP06]

Equivalence Class[LWFP06] A table's equivalence class for an attribute set is the collection of all tuples in the table that have the same values for the attribute set.

For example, tuples 1 and 2 in Figure 2.1(a) form an equivalence class concerning attribute set Gender, Age, Postcode . Their corresponding values are identical.

k-anonymity Property[LWFP06]: A table is k-anonymous with respect to a quasi-identifier if the size of every equivalence class is k or more.

k-anonymity requires that every occurrence within an attribute set has a frequency of at least k. For example, Figure 2.1(a) does not satisfy 2-anonymity property since tuples {male, young, 4351} and {female, young, 4352} occur only once.

The original data must be transformed using data anonymization technology in order to achieve the goal of data anonymization. The process of replacing the original value with a vague, ranged value which is called **generalization** offers the broadest variety of applications and can represent the properties of the original data set well.

Generalization may be achieved in two ways: global recoding and local recoding. **Global recoding** takes place at the domain level. Every instance of an attribute value is replaced with a new generalized value when an attribute value is generalized. A table could be overgeneralized using a global recoding approach. Figure 2.1(c) provides a global recording example.

A **local recoding** method generalizes attribute values at the tuple level. The original value of the attribute coexists alongside a generalized value. An anonymous view's distortion could be minimized by a local recoding approach since it does not overgeneralize a table. However, optimal local recoding is NP-hard. An example of local recoding is given in Figure 2.1(b).

2.1.2 L-Diversity

L-diversity:[MKGV07] An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity.

Well-represented: The simplest understanding of "well represented" would be to ensure there are at least l distinct values for the sensitive attribute in each equivalence class. Under this definition, l-diversity is called probabilistic l-diversity.

According to Machanavajjhala's study[MKGV07], there are two more definitions of "well-represented" that can protect against attacks using probabilistic inference, namely entropy l-diversity and recursive (c, l)-diversity. In this research, we only use probabilistic l-diversity.

2.1.3 T-Closeness

t-closeness[LLV06]: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

As can be shown, defining the distance between two distributions is important for determining t-closeness. The original research uses the Earth Mover Distance (EMD). EMD has the advantage that it considers the semantic relationship between attributes. We also use EMD in this research. The calculation of EMD is omitted here.

2.2 Problem Description

2.2.1 Data Disclosure

It's necessary to secure against the disclosure of personal sensitive information while publishing data. In the literature[DL86], there are two different types of information disclosure:

(1) **Identity disclosure** occurs if a third party can identify a subject or respondent from the released data. For instance, a data table includes information about a woman who has 18 children. Due to the rareness of women with 18 children, this woman's private information can be discovered by combining attributes from publically accessible data sources (e.g. voter registration data).

(2) **Attribute disclosure** occurs when sensitive information about a person or the workings of a facility is exposed or can be accurately predicted. For example, tabular data shows hospital parameters including patient counts and average age. It can be assumed that a hospital is a children's hospital if it does not treat patients older than 18 years. Therefore, this attribute could help to learn something about the facility.

Attribute disclosure frequently follows the disclosure of identity. An individual is reidentified after identity revelation, and the associated sensitive values are revealed to the public. Identity disclosure is not a need for attribute disclosure to take place.

2.2.2 Data Attack

K-anonymity offers straightforward and clear protection. Anyone who knows just the quasi-identifier values of one individual cannot identify the record belonging to that individual with confidence greater than 1/k if the table fulfills k-anonymity for some value k. K-anonymity offers some security against attribute disclosure but not enough against identity disclosure. Several writers have acknowledged this, including[MKGV07][XT06]. The homogeneity attack and the background knowledge attack were the two attacks named in[MKGV07]. We take Figure 2.2 as an example. (a) is the original data, and (b) is a 3-anonymized version. The Disease attribute is sensitive.

(1) **Homogeneity attack**[LLV06]: Assume Alice knows that Bob is a 27-year-old man living in ZIP 47678 and Bob's record is in the table. From

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Figure 2.2: (a) left: Original data (b) right: 3-anonymous data[LLV06]

(b), Alice can conclude that Bob corresponds to one of the first three records and thus must have heart disease.

(2) **Background knowledge attack**[LLV06]: Assume that Alice can determine that Carl belongs to a record in the final equivalence class in (b) based on Carl's age and zip code. Furthermore, suppose that Alice is aware of Carl's extremely low chance of having heart disease. With this background information, Alice can determine that Carl most likely has cancer.

It can be deduced that l-diversity can protect data from attribute disclosure by avoiding homogeneity attacks. However, it has also some disadvantages.

(3) **Similarity Attack**[LLV06]: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. Consider the following example.

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

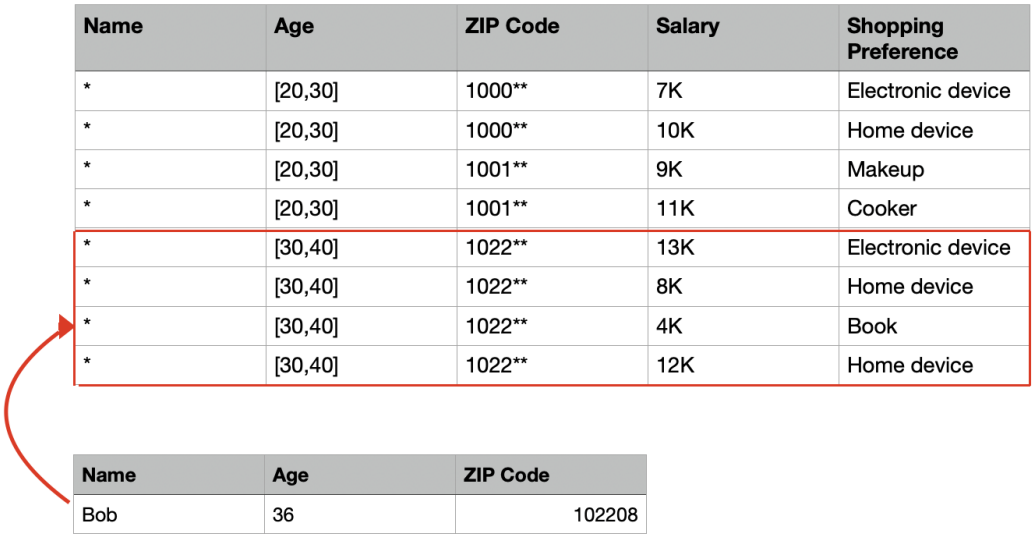
Figure 2.3: (a) left: Original data (b) right: 3-diverse data[LLV06]

The original table is shown in Figure 2.3(a), while Figure 2.3(b) displays an anonymized version meeting 3 distinct diversity requirements. Salary and Disease are the two sensitive attributes. If Alice knows that Bob's record matches one of the first three records, she will know that Bob's pay is between [3K-5K] and will be able to deduce that it is a tiny amount (she does not care about

the exact value). For categorical attributes such as "Disease", because all three of the diseases in the group are stomach-related, Alice knows that Bob may have some stomach-related issues.

While the l-diversity criterion provides "diversity" of sensitive values in each group, it does not account for the semantic similarity of these values, which leads to the leaking of sensitive information.

Finally, the privacy will still be leaked if the data satisfies all the k-anonymity, l-diversity, and t-closeness. Given multiple sensitive attributes, background attacks against sensitive attributes will still work. As the instance shows in Figure 2.4, salary and shopping preference are sensitive attributes. Alice knows Bob's age and ZIP code, and she also knows that Bob has many books. It's easy to find Bob's salary. So no matter what anonymity method is applied, privacy leakage is still difficult to avoid. But this situation rarely happens, so we think that most privacy leakage problems can be solved through k-anonymity, l-diversity, and t-closeness.



Name	Age	ZIP Code	Salary	Shopping Preference
*	[20,30]	1000**	7K	Electronic device
*	[20,30]	1000**	10K	Home device
*	[20,30]	1001**	9K	Makeup
*	[20,30]	1001**	11K	Cooker
*	[30,40]	1022**	13K	Electronic device
*	[30,40]	1022**	8K	Home device
*	[30,40]	1022**	4K	Book
*	[30,40]	1022**	12K	Home device

Name	Age	ZIP Code
Bob	36	102208

Figure 2.4: 2-anonymity, 2-diversity, t-closeness

2.3 Usability Testing Questionnaires

As the field of information technology advances rapidly, there is a growing emphasis on understanding and improving user experience. In the realm of user interface design, usability testing questionnaires have become essential tools for evaluating and enhancing system interaction performance. This section introduces several classic usability testing questionnaires, involving their merits, drawbacks, and suitable scenarios.

2.3.1 System Usability Scale

The System Usability Scale (SUS), proposed by John Brooke in 1986[Bro86], stands as a widely employed questionnaire. Using a series of brief statements and ratings, SUS assesses users' subjective perceptions of the overall usability of a system. Its advantages include simplicity and broad applicability across various systems, while its main limitation lies in the subjectivity of the obtained scores. SUS consists of a set of ten questions, each rated on a Likert scale from 1 to 5, where 1 represents "strongly disagree" and 5 indicates "strongly agree." The questions are carefully crafted to cover diverse aspects of the user experience, including system complexity, ease of use, and overall satisfaction.

1. I think that I would like to use this system frequently if I have the demand.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I think there was too much inconsistency in this system.
7. I think the feedback and tips in this system is clear.
8. I found the system very cumbersome to use.
9. I think the difficulty of using the various functions in this system is similar.
10. I needed to learn a lot of things before I could get going with this system.

The scoring methodology involves converting the individual Likert scale responses into numerical values, and then applying a specific formula to derive a usability score that ranges from 0 to 100. Higher scores indicate a more favorable user perception of usability. Overall usability is reflected in SUS scores. Researchers (Bangor, 2009[BKM09]) have determined the correlation between words, letters, allowable ranges, and SUS scores for overall usability. The meaning of SUS scores may be seen and clarified visually in Figure 2.5.

One of SUS's notable strengths lies in its simplicity. The straightforward questionnaire design allows for quick and efficient administration to a large number of participants. Moreover, SUS is applicable across a broad spectrum of systems, ranging from software interfaces to websites and hardware devices. Its adaptability contributes to its widespread use in diverse industries. Furthermore, the standardized nature of SUS facilitates benchmarking

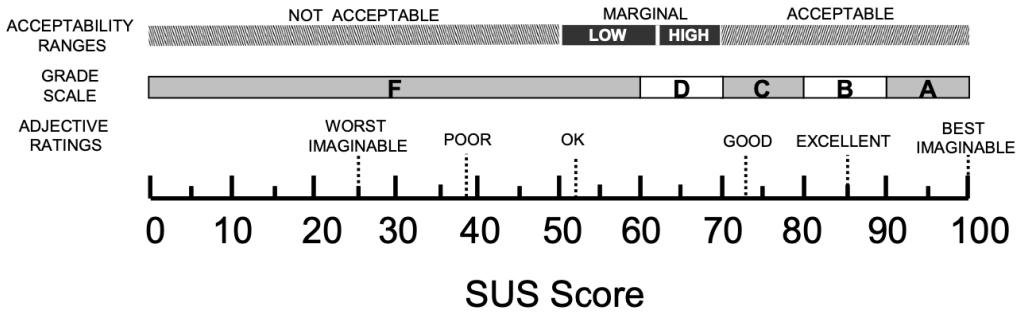


Figure 2.5: A comparison of the adjective ratings, acceptability scores, and school grading scales, in relation to the average SUS score[BKM09]

and comparison across different studies, enabling researchers to gauge usability improvements over time.

However, its limitation is also obvious. As with any self-reported questionnaire, SUS is subject to individual interpretation. Users' responses may be influenced by personal preferences and experiences. Besides, while effective for obtaining an overall usability score, SUS might not provide detailed insights into specific usability issues or areas for improvement.

The application of SUS has been documented in numerous case studies across various domains, including e-commerce platforms[LS09], healthcare systems[BKM08], and mobile applications[Sau11]. Researchers often combine SUS with other usability metrics to gain a comprehensive understanding of the user experience. In conclusion, the System Usability Scale continues to be a valuable tool for usability tests, offering a balance between simplicity and effectiveness in evaluating the perceived usability of systems.

2.3.2 After Scenarios Questionnaire

The After Scenarios Questionnaire (ASQ)[Lew91] is a user feedback tool designed to assess the user experience immediately after participants engage in specific scenarios or tasks within a system. Developed as part of the Usability Engineering Process by the Software Usability Research Laboratory (SURL) at Wichita State University. The ASQ typically includes questions related to the user's perception of the difficulty, efficiency, and satisfaction of completing tasks. Here is a general overview of the types of questions found in the ASQ:

1. **Difficulty Rating:** Overall, I am satisfied with the ease of completing the task in this scenario.

2. **Efficiency:** Overall, I am satisfied with the amount of time it took to complete the task in this scenario.

3. **Satisfaction:** Overall, I am satisfied with the support information (tips, user interaction) when completing the task.

Scoring is often done on a Likert scale, where respondents rate each item on a numerical scale (e.g., 1 to 7), and the scores are then averaged to provide an overall measure of usability. Lower scores on difficulty and higher scores on efficiency and satisfaction are generally indicative of better usability.

ASQ has been recognized for its effectiveness in offering detailed insights into users' experiences with specific tasks or scenarios, facilitating a nuanced evaluation. Administered directly after task completion, ASQ captures users' fresh impressions, minimizing recall bias and ensuring timely insights. Moreover, ASQ's focus on scenarios aligns with real-world user interactions, offering context-specific feedback relevant to the actual system usage. However, ASQ is best suited for evaluations tied to specific scenarios, and its application may be less effective for broader system-wide assessments. Besides, the immediate nature of feedback may introduce emotional bias, influenced by participants' temporary feelings following task completion. In conclusion, ASQ serves as a valuable tool for evaluating user experiences within the context of specific tasks or scenarios. Its immediate feedback and task-specific focus make it an effective instrument for understanding user perceptions in targeted usability assessments.

2.3.3 Confidence Rate

The Confidence Rate (CR) is a metric used in usability testing to assess the level of confidence users have in completing specific tasks or scenarios. It is often employed as a subjective measure to capture users' self-reported confidence in their ability to perform a task successfully. The CR is typically determined through self-assessment by the users themselves.

Users are asked to rate their confidence in completing a task or scenario after they have attempted it. The confidence rating is usually measured on a numerical scale or a Likert scale. CR is not explicitly defined or standardized across all studies. Here we give the question of CR in this study:

- Are you sure that you have done a good decision to balance the utility and privacy for anonymized data?

The question is designed to gauge users' confidence in their decision-making regarding the trade-off between utility and privacy in anonymized data. The wording emphasizes certainty, prompting users to express their level of assurance. This approach aims to capture nuanced responses and provide insights into users' confidence in their choices.

It's important to note that the Confidence Rate is a self-reported measure and reflects users' perceptions, which may not always align perfectly with objective task performance. However, it can still provide valuable insights into users' comfort levels and perceptions during usability testing.

2.3.4 Expectation Rate

Expectation Rate (ER) is not a commonly recognized or standardized term in the field of usability testing or user experience research. The concept of Expectation Rate proposed by Albert and Dixon (2003)[AD03], involves users providing subjective ratings or expectations regarding the anticipated difficulty of tasks before they perform them. This measure is then compared with their actual experiences during or after task completion. Both the expectation and the reality about performing a given task are ranked on a scale from (1) very easy to (7) very difficult. With this, after performing the test it is possible to construct a matrix divided into four quadrants, as shown in the Figure 2.6:

- **Don't touch it**: Tasks that users rated as easy before and after their execution. So it is feasible that no action is taken in relation to them.

- **Fix it fast**: Tasks that the participants thought were easy, but turned out difficult during the test, which is nothing more than an indicator of high user dissatisfaction. This is the quadrant with the highest priority to make changes.

- **Big opportunity**: Tasks that users defined as difficult before and after their accomplishment. This is a good opportunity to make improvements and move the tasks to the "Promote it" quadrant.

- **Promote it**: Tasks that users had the perception of being difficult before execution, but were classified as easy after it. At this point, it is worth reflecting on how to promote this functionality.

Expectation Rate is used to assess users' perceptions of task difficulty before engaging with a system or interface. This can provide insights into users' preconceived notions and expectations, allowing researchers to analyze how closely these expectations align with the actual usability of the system. Researchers might use the Expectation Rate as part of a broader methodology to understand users' subjective experiences, combining it with other usability metrics or qualitative feedback to form a comprehensive evaluation.

2.4 Methodology

2.4.1 Selection of Datasets

The selection of appropriate datasets plays a pivotal role in the robustness and relevance of our case study. The criteria for choosing datasets are guided by a careful balance between representativeness, diversity, and alignment with the objectives of the study. We consider the following perspectives by selecting datasets:

Figure 2.6: Expectation Rate[AD03]

1. **Real-world Relevance**. The datasets chosen mirror real-world scenarios and applications relevant to the intended use of the developed prototype. Emphasis is placed on datasets derived from domains where the balance between utility and privacy is a critical consideration.

2. **Diversity of Data Structures**. Datasets with diverse structures are included. This encompasses variations in the number of features, and data types, allowing for a comprehensive evaluation across different use cases.

3. **Privacy Sensitivity**. Datasets are selected to encompass a spectrum of privacy sensitivity. This includes datasets containing both sensitive and non-sensitive attributes, providing a nuanced understanding of the prototype's efficacy in handling varying degrees of privacy concerns.

4. **Scalability**. The scalability of the prototype is assessed through datasets of varying sizes. This consideration is essential to evaluate the system's performance across datasets of different scales, providing insights into its adaptability to diverse data volumes.

By adhering to these considerations, our dataset selection aims to create a comprehensive and rigorous evaluation environment for the prototype. The diverse nature of the chosen datasets ensures a thorough examination of the system's capabilities, allowing us to draw meaningful conclusions regarding its effectiveness and adaptability across different practical scenarios.

We use the data sets from Kaggle[kag], which offers a strategic advantage for research endeavors. Kaggle, a prominent platform in the data science community, provides researchers with diverse datasets spanning various domains. This diversity ensures that researchers can find datasets aligned with their specific research objectives. Moreover, Kaggle datasets undergo meticulous quality control, ensuring a high standard of reliability. The accompanying documentation and descriptions provide valuable insights into data sources, structures, and features, enhancing researchers' understanding of the data. The Kaggle community further strengthens this choice, offering a collaborative space for knowledge exchange, discussions, and feedback. This supportive environment fosters a rich learning experience for researchers. Additionally, Kaggle datasets often reflect real-world challenges, making them applicable to practical scenarios through competitions and tasks. Leveraging these datasets allows researchers to directly apply their findings to real-world problem-solving, enhancing the practical relevance of their work. In essence, the choice of Kaggle datasets provides researchers with a diverse, reliable, and well-supported foundation, fostering comprehensive exploration within the dynamic Kaggle community.

2.4.2 Sample Size Determination

In establishing the sample size for our usability testing, we turn to Nielsen's influential "five users is enough" formula, as outlined in a 1993 paper co-authored with Tom Landauer[NL93]. This formula, rooted in statistical principles, guides our approach to determining optimal sample size. Leveraging the binomial probability theorem, allows us to estimate the probability of encountering usability issues within the user population. We use a tool based on Nielsen's formula to calculate the sample size of our usability test [mea]. This calculator produces an estimate of the number of users needed to discover the specified percent of total problems. In our usability testing, our goal is to uncover approximately 95% of potential issues. To identify even the less frequent challenges, we assume a conservative probability of 25% for the occurrence of usability issues within the user population. While Nielsen's research suggests a reasonable estimate of 30%, we opt for a more stringent approach to enhance the reliability of our testing. The calculated result Figure 2.7 shows that our sample size should be on average 10.41 users. In the usability test, we recruited 10 users in each iteration.

2.4.3 Combining Quantitative and Qualitative Approaches in Usability Test

In the usability test, we adopt a comprehensive approach by integrating both quantitative and qualitative methods to gain a holistic understanding of user experience. This dual-method strategy allows us to leverage the strengths of each approach and provide a more nuanced evaluation.

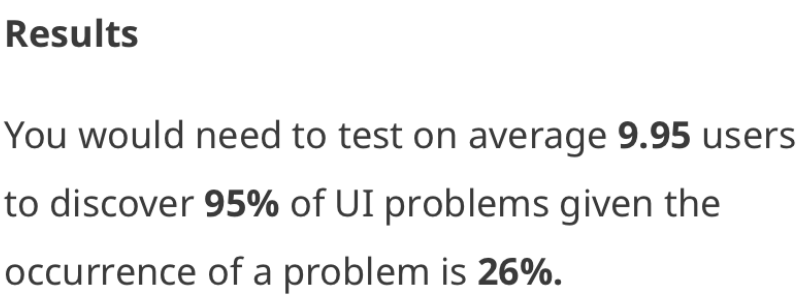


Figure 2.7: Sample Size calculated by [mea]

For the quantitative aspect, we employ survey questionnaires distributed to participants involved in comparative experiments. The sample size comprises a total of 30 participants, with 10 individuals participating in each usability analysis session. While this sample size may fall short of the ideal quantity for rigorous quantitative analysis, it serves as a robust foundation for outlining general trends and patterns in usability metrics. In tandem with quantitative data collection, we employ qualitative methods through remote video conferences with users. Using the "think-aloud" method, participants are encouraged to verbalize their thoughts and actions during interactions with the prototype. This qualitative approach allows us to delve deeper into the users' cognitive processes, revealing insights into their decision-making and problem-solving strategies. Through these sessions, we identify and analyze instances where users encounter challenges, misunderstand functionalities, or express preferences.

In the context of usability testing with a relatively small sample size, the integration of qualitative and quantitative analysis proves to be a practical and insightful approach. While traditional usability studies often emphasize large sample sizes for statistical robustness, several considerations make the combination of qualitative and quantitative methods valuable in smaller-scale studies. First, qualitative analysis, particularly through methods like "think-aloud" sessions and in-depth interviews, provides a rich understanding of users' behaviors, attitudes, and challenges. In a small sample, qualitative insights can offer depth and context that quantitative metrics alone may lack, contributing to a more nuanced interpretation of user experiences. Second, in small-scale studies, individual user experiences play a more significant role. Qualitative analysis allows researchers to identify specific usability issues, uncover user misconceptions, and gain insights into the reasons behind certain behaviors. This qualitative depth can be instrumental in refining the user interface and addressing issues that might be overlooked in a larger sample. Third, quantitative metrics, even with a small sample, provide measurable data on usability performance. While statistical significance may be challenging to achieve with a limited sample size, quantitative measures still offer valuable benchmarks and trends. Finally, the iterative nature of qualitative feedback becomes crucial in

small sample studies. Qualitative insights guide iterative design improvements, allowing researchers to quickly implement changes and observe their impact. This iterative process is conducive to refining the prototype, addressing user concerns, and enhancing overall usability.

2.5 Related Work

Numerous privacy-preserving data visualizations exist, but few prioritize the crucial task of balancing the trade-off between privacy and utility. Wang's work, such as [WCC+17], aids data controllers in understanding how privacy parameters impact data transformation, leading to visualizations striking a balance between privacy gain and utility loss. Chou[CWM16] proposes a visual interface identifying potential privacy concerns and maintaining data utility in event sequence data analysis. Studies like [MKGV07], although lacking visualization interfaces, inspire our measurement of the trade-off. These studies employ various measurement methods for utility and privacy, as well as distinct visualization techniques.

Chou[CWM16] integrates anonymization methods into visualization, introducing privacy-preserving operations like Merge, Redirect, Delete, and Blur. However, quantitative evaluation methods for privacy and utility are lacking. Wang[WCC+17] presents a tool visualizing utility and privacy, offering feedback on utility changes during anonymization. Wang's focus is on finding privacy disclosure, analyzing utility and privacy separately. Li[LL09] proposes an integrated framework considering the privacy-utility trade-off, borrowing from the Modern Portfolio Theory. Our study incorporates Li's concept to provide data controllers with guidelines for the right trade-off.

Various methods measure utility and privacy, like Generalization Height[LDR05] and the discernibility metric[BA05]. However, they neglect data distribution. Machanavajjhala[MKGV07] uses KL-Divergence to describe the distance between anonymized and original data, presenting information loss.

In our study, we concur with Chou[CWM16], recognizing diverse user privacy considerations. Providing users flexibility to fine-tune anonymization results achieves a better utility-privacy trade-off. Wang's[WCC+17] tool highlights how good visualizations offer an intuitive understanding of utility and privacy changes during anonymization. Applying quantitation methods boosts the tool's confidence. We use distribution distance, as done by Machanavajjhala[MKGV07]. Here we summarize the features of related work ff in Table 2.1. These studies applied different measurement methods for utility and privacy, as well as different visualization techniques.

Study	U. Vis	P. Vis	Measurement	Vis Tech
Chou[CWM16]				Sankey Diagram
Wang[WCC+17]	✓	✓	Generalization Height	Matrix, Tree
LeFevre[LDR05]			Generalization Height	
Bayardo[BA05]			Discernibility Metric	Matrix
Macha[MKGV07]			KL-Divergence	
Li[LL09]	✓	✓	KL-Divergence	Scatter Plot

Table 2.1: Features of privacy-preserving systems supporting evaluating trade-offs." U. Vis" indicates "Utility visualization" and "P. Vis" indicates "Privacy visualization".

3 Implementation

This chapter delineates the practical realization of our prototype, focusing on three core aspects. First, we delve into implementing anonymization methods, elucidating the Mondrian algorithm employed to safeguard sensitive information. This section provides an in-depth exploration of the technical underpinnings that form the foundation of our privacy-preserving approach. Moving forward, we turn our attention to the quantification of the utility and privacy levels of anonymized data. Methodologies for evaluating the effectiveness of anonymization techniques in balancing privacy and utility are expounded upon. Additionally, the chapter unfolds the implementation of the application prototype. Comprehensive insights into the architectural decisions, chosen frameworks, and the iterative development of low-fidelity and high-fidelity prototypes are discussed. By dissecting the intricate threads of algorithmic implementation, privacy quantification, and application prototyping, this chapter lays the groundwork for a nuanced comprehension of the practical realization of our proposed solution.

3.1 Algorithms

3.1.1 Mondrian for data anonymization

We use the Mondrian algorithm to achieve k-anonymity, l-diversity, and t-closeness. Python is the programming language, and libraries used include Pandas and Numpy.

According to publications written by Kristen LeFevre[LDR06], Mondrian is a Top-down greedy data anonymization technique for relational datasets. Mondrian is a fast local recording method. It also maintains high data utility. The original Mondrian proposed by [LDR06] only achieves k-anonymity. In this research, we also add l-diversity and t-closeness in the process, which is implemented by Telesoho[Tel].

The basic workflow of Mondrian is:

(1) Partition the raw dataset into k-groups using kd-tree. k-groups means that each group contains at least k records.

(2) Generalization (Figure 2.1.1) each k-group, such that each group has the same QI(Section 2.1).

In the first step, we use kd-tree to partition the data. Kd-tree is a fast, straightforward, and sufficient k-dimensional binary tree. Each node of the

Figure 3.1: Workflow of basic Mondrian

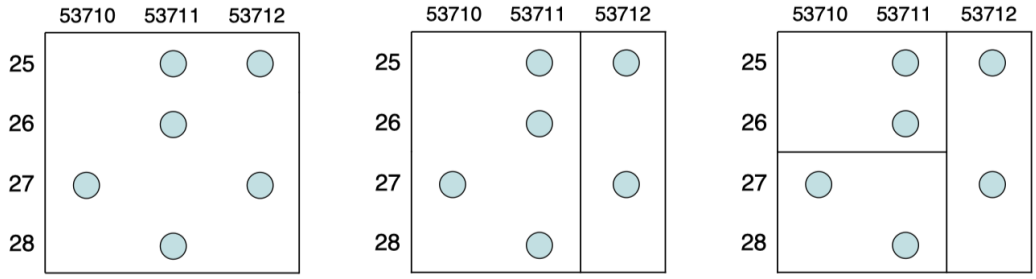


Figure 3.2: Spatial representation of Patients and partitionings (quasi-identifiers Zipcode and Age)[LDR06]

tree is k-dimensional. Figure 3.2 shows the simple partition (k=2) process of a 2-dimensional tree.

A top-down greedy algorithm for multidimensional partitioning is displayed below. ϕ denote local recoding functions, maps each tuple $t \in T$ to some recoded tuple. The dimension and value to split about must be selected for each iteration. Median partitioning was one method mentioned in the kd-tree literature[FBF77] for achieving uniform occupancy. The median of the partition projected on dim is used as the split value. The time complexity is O(nlogn) where n=|T|, the same as when building a kd-tree.

Algorithm 1 anonymize(*partition*)[LDR06]

if (no allowable multi-dimensional cut for *partition*) **then**
  return $\phi : partition \rightarrow summary$
else
  $dim \leftarrow$ choose_dimension()
  $fs \leftarrow$frequency_set(*partition*,*dim*)
  $splitVal \leftarrow$find_median(fs)
  $lhs \leftarrow \{t \in partition : t.dim \leq splitVal\}$
  $rhs \leftarrow \{t \in partition : t.dim > splitVal\}$
  return Anonymize(rhs)$\cup$Anonymize(lhs)
end if

To let the basic Mondrian algorithm fit the requirements of l-diversity and t-closeness, we just need to add a check step before partitioning.

3.1.2 Privacy Criteria Control

Not all data controllers are very familiar with anonymization algorithms. According to the definition of k-anonymity, l-diversity, and t-closeness, we can understand the meaning of the parameters k, l, and t, and know that if we want to achieve higher privacy requirements, we need to assign larger k, l values, and small t value. The main issue is that these three parameters do not vary on the same scale. For example, t is normally between 0.0 and 1.0, but k and l are integers. Furthermore, because the number of records and the number of various sensitive attribute values are on different scales, setting k and l to the same value is not practical. To ease privacy control, the study employs a uniform parameter p introduced by J.K. Chou[CWM16]. The parameter p has a range of 0.0 to 1.0 and combines the three privacy requirements into one single intuitive value:

$$k = max(1, \lceil p * k_{max} \rceil) \tag{3.1}$$

$$l = max(1, \lceil log_2 k \rceil) \tag{3.2}$$

$$t = max(t_{min}, \frac{l_{max}}{1 + (l * p)} * t_{min}) \tag{3.3}$$

k_{max} and l_{max} are the maximum allowed k and l values respectively. t_{min} refers to the minimum allowed t value. In the experiment of this research, the value of $k_{max}, l_{max}, t_{min}$ should be settled first. k_{max} and l_{max} can be settled automatically according to the value domain of attributes, and t_{min} is usually 0.25 according to the research of J.K. Chou[CWM16]. However, [CWM16] did not give the basis for settling these parameters. We will discuss this in Section 4.1. The goal of introducing p is to provide the data controllers with a clearer way to adjust their preferred privacy level, with a smaller p value indicating a relatively loose privacy level and a larger p value indicating a higher privacy need.

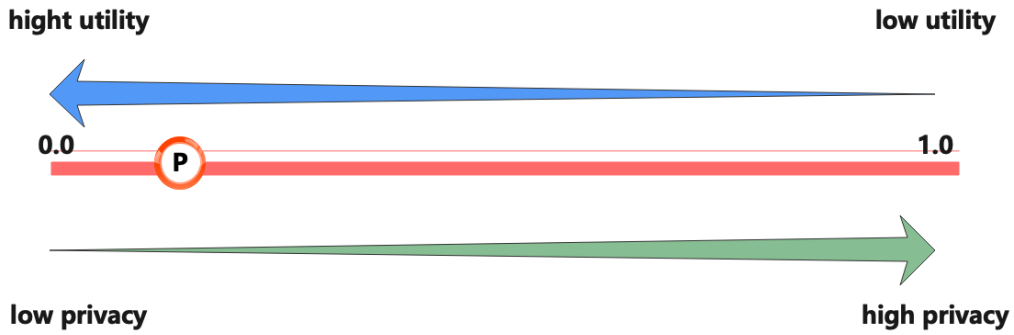


Figure 3.3: adjust parameter p to get the best visualization

3.2 Quantification of Utility and Privacy

To measure utility and privacy, several characteristics of utility and privacy are given. Li [LL09] identified three characteristics as follows:

(1) Specific knowledge (that about a small group of individuals) has a larger impact on privacy, while aggregate information (that about a large group of individuals) has a larger impact on utility;

(2) Privacy is an individual concept and should be measured separately for every individual while utility is an aggregate concept and should be measured accumulatively for all useful knowledge.

(3) Any information that deviates from the prior belief, false or correct, may cause privacy loss but only correct information contributes to utility.

The following implications on measuring privacy and utility are implied by these three characteristics[LL09]:

(1) For privacy, the worst-case privacy loss should be measured. For utility, the aggregated utility should be measured.

(2) Privacy should be measured against the trivially anonymized data whereas utility should be measured using the original data as the baseline.

3.2.1 Measuring Privacy Loss

The implication above is an indicator for measuring privacy and utility. [LL09] proposed a worst-case privacy measure, which introduced entropy to describe the "Distance" between distributions before and after anonymization. Let Q be the distribution of the sensitive attribute (SA) over the whole data set. Since the data attacker always has access to the distribution Q, even if all quasi-identifiers (QIs) are generalized, we take Q as the prior knowledge of the data. Only when the data attacker learns private information from outside of distribution Q, do privacy leaks occur. The posterior knowledge about a tuple's sensitive attribute s shrinks to the equivalence class that contains s when the data attacker gets the anonymized data. Let $P(s)$ represent the equivalence class's distribution of the sensitive attribute. The distance between Q and $P(s)$ is used to calculate the privacy loss for a tuple t. The JS-divergence distance metric is engaged:

$$P_{loss}(s) = JS(Q, P(s)) = \frac{1}{2}[KL(Q, M) + KL(P(s), M)] \tag{3.4}$$

Where $M = \frac{1}{2}(Q + P(s))$ and $KL()$ is KL-divergence[KL51].

The worst-case privacy loss is the maximum privacy loss of all the tuples in the data:

$$P_{loss} = max(P(s)) \tag{3.5}$$

JS-divergence is just one of the methods of measuring distance. Other measurements can also be chosen.

3.2.2 Measuring Utility Loss

There are two main approaches to measuring utility. One measures the amount of utility that remains in the anonymized data. This includes the sum[XWP+06] or the average[LDR05] of the interval size in the equivalence class. The other approach measures the utility by comparing the anonymized data with the original data. This includes measures such as the number of generalization steps and the KL-divergence between the restricted distribution and the original distribution for all possible quasi-identifier values[MKGV07]. According to the argument of [LL09], the utility should be measured as utility loss against the original data. Therefore this study adopts the second approach by applying entropy (for instance KL-divergence and JS-divergence) to describe the distance between distributions before and after anonymization. According to [LL09], the main task is to find the distribution of sensitive attributes for large populations. A large population can be specified by a support value and a predicate involving only quasi-identifiers. For instance a set {tuple|tuple.Age >=44 && tuple.Sex = Male}. The support value is the number of records that satisfy the predicate. [LL09] proposed a method for measuring utility loss that is used in this research.

First, we find all large populations whose support values are at least a threshold value defined by users (data controller in this research). The predicates involved not only the values from the attribute's domain but also from the generalized hierarchy of the quasi-identifiers (e.g. Figure 3.4 shows the generalized hierarchy of a quasi-identifier.) We use the FP-tree algorithm[HPY00] for finding all large populations. Next step, we calculate the estimated distribution $\overline{P}_y$ of each large population y of the sensitive attribute from the anonymized data and the true distribution P_y of the sensitive attribute from the original data. Again, we use the JS-divergence as the distance measure between $\overline{P}_y$ and P_y to describe the information loss from the view of the data consumer. Finally, according to the argument of [LL09], utility is an aggregate concept, we use the average utility loss U_{loss} of all large populations.

$$U_{loss}(y) = JS(P_y, \overline{P}_y) \tag{3.6}$$

$$U_{loss} = \frac{1}{Y} \sum_{y \in Y} U_{loss}(y) \tag{3.7}$$

where Y is the set of all large populations. The anonymized data provides the maximum utility when $U_{loss} = 0$.

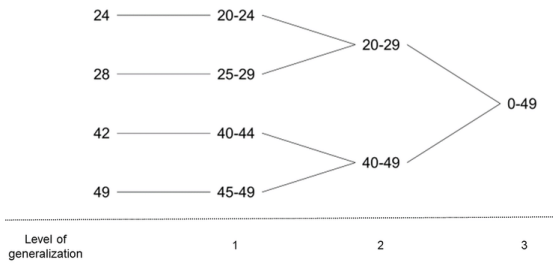


Figure 3.4: Generalization hierarchy of a quasi-identifier attribute[DPCTH22]

3.2.3 Measuring Trade-off

We describe how to quantify utility and privacy separately in Section 3.2.1 and Section 3.2.2. However, there is no relevant work for quantifying the privacy and utility trade-off as an entirety. In this research, we are trying to give a method of measuring the trade-off between privacy and utility. We define the effect of trade-off as follows:

$$E_{trade-off} = g(P_{loss}, U_{loss}) \tag{3.8}$$

where P_{loss} and U_{loss} indicate privacy loss and utility loss respectively. $g()$ is an undefined operation. Since P_{loss} and U_{loss} are negative correlated, we use power function $f(x) = x^a, a \in (-\infty, 0]$ to fit the correlation between P_{loss} and U_{loss}. We assume that privacy and utility have the same weight when we consider the trade-off so we set $a = -1$ (Of course, the value a can also take other values to adjust the weights of U_loss and P_loss. This does not affect the overall effect of the model. We only set $a = -1$ to make the model simple.) Therefore we have:

$$U_{loss} = f(P_{loss}) = c * (P_{loss})^{-1} \tag{3.9}$$

Where c is a constant value. Since both P_{loss} and $U_{loss} \in [0, 1]$, c would be a small value if both P_{loss} and U_{loss} are small. So we have:

$$c \propto E_{trade-off} \tag{3.10}$$

We add the correction factor $(P_{loss} - U_{loss})^2$ to calibrate the value of $E_{trade-off}$ in extreme cases (e.g. $U_{loss} = 0$ and $P_{loss} = 1$). So according to Equation 3.10 and Equation 3.9 we get:

$$E_{trade-off} = g(P_{loss}, U_{loss}) = (P_{loss} * U_{loss} + (P_{loss} - U_{loss})^2)^{-1} \tag{3.11}$$

Figure 3.5 provides an intuitive understanding of the nature of $E_{trade-off}$, we observe that Equation 3.11 exhibits a clear monotonic trend. As both P_{loss}

and U_{loss} decrease, the value $E_{trade-off}$ demonstrates a gradual rise. Notably, the plot attains a maximum value in the vicinity of the origin, aligning with our expectations. This characteristic allows $E_{trade-off}$ to effectively capture the trade-off between utility and privacy. In this region, we achieve a certain optimal trade-off. Furthermore, the plot reveals smooth and monotonic behavior in the area near the P-axis and U-axis. This implies that in these regions, further adjustments to the parameters may not significantly enhance the $E_{trade-off}$ value. This observation aligns with our considerations regarding the trade-off between utility and privacy in the vicinity of extreme values, suggesting that additional parameter tuning may not yield substantially improved results in certain scenarios.

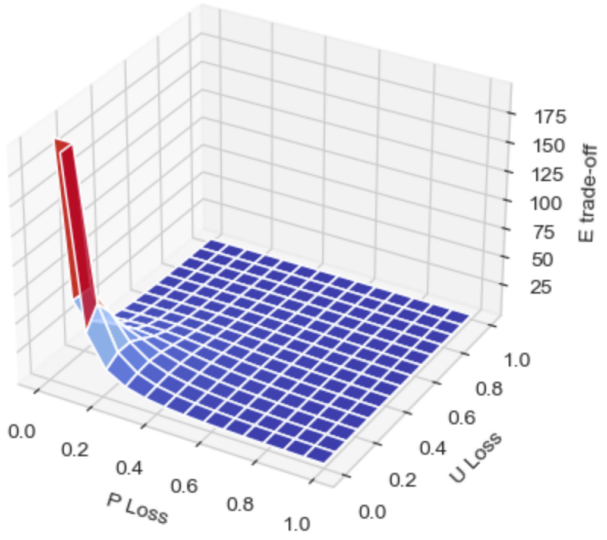


Figure 3.5: 3D Chart of $E_{trade-off}$

3.3 Prototyping

In this chapter, we delve into the process of designing and implementing prototypes, spanning from low-fidelity to high-fidelity iterations. This chapter serves as a comprehensive exploration of the prototyping phase, encompassing key aspects such as design considerations, deployment strategies, and the development environment.

The journey begins with the creation of low-fidelity prototypes, which provide a preliminary visualization of the proposed system. These prototypes, which are embedded in the first usability test (Section 4.2.1), though simple in appearance, serve as invaluable tools for gathering early feedback and refining the conceptual design. We discuss the strategic choices made in the design process, emphasizing user experience and functionality while keeping development costs at a minimum.

Transitioning to high-fidelity prototypes, we explore the refinement of design elements and the incorporation of intricate details that mirror the envisioned final product after gathering feedback from the usability test (Section 4.2.1). This phase involves a more sophisticated approach to user interface design, ensuring a seamless and intuitive interaction.

The chapter aims to provide a holistic view of prototyping, offering readers a nuanced understanding of the decisions and considerations that underpin the development of both low-fidelity and high-fidelity prototypes.

3.3.1 Low Fidelity Prototype

Development Framework

In the development of the low-fidelity prototype, Python served as the primary programming language, leveraging the streamlined capabilities of the Streamlit framework[str] for rapid interface development. The decision to utilize Streamlit was driven by its intuitive nature, allowing for the creation of user-friendly interfaces without the need for an extensive background in front-end development. The framework's out-of-the-box UI components [JWX10] and basic interactivity features facilitated a swift and effective prototyping process, aligning seamlessly with the project's emphasis on quick iteration.

For data manipulation and processing, the prototype heavily relied on the Pandas library [Pan], capitalizing on its efficiency in handling microdata. Pandas provided a robust foundation for implementing algorithms for conducting data analysis, offering a convenient and expressive syntax that expedited the integration of complex data operations into the prototype.

The choice of PyCharm as the integrated development environment (IDE) further streamlined the coding process, providing a comprehensive set of tools and features for Python development. This IDE selection ensured a cohesive and efficient development workflow, enhancing code readability and maintainability.

In essence, the low-fidelity prototype development phase was characterized by a strategic alignment of tools and frameworks. This approach aimed to accelerate the prototyping cycle, allowing for swift adjustments based on user feedback and ensuring the timely evolution of the software solution.

UI Design

In order to compare the traditional approach and our approach of anonymization selection, we conduct multiple rounds of experiments to identify UI components and business logic combinations with higher usability. Traditional heuristic approaches rely on user experience and the results of anonymized data to assess whether it meets privacy and utility requirements. Hence, we integrated a **spreadsheet** as a baseline into the low-fidelity prototype. Additionally, by applying the quantification of privacy and utility from Section 3.2,

we incorporated **numerical indicators** into the prototype to present thew utility loss and privacy loss of anonymized data seperately, independent of visualization support. To further explore and compare the usability, we included a **scatter plot** as our proposed visualization method, aiming to provide users with an intuitive understanding of the trade-off between privacy and utility during anonymization. On the business logic side, we integrated Chou's **PCC** method[CWM16] and our **auto PCC** method into the prototype for comparative usability testing.

In terms of basic functionality, the low-fidelity prototype ensures adherence to the Mondrian algorithm's requirements. Users are prompted to input their chosen anonymization method and parameters (excluding auto PCC, which does not require user input). Users are also able to select Quasi-Identifiers (QIs) and Sensitive Attributes (SAs). Additionally, we provided a data uploader for users to upload raw data and a data downloader for retrieving anonymized data. Table 3.1 illustrates the fundamental comparison objects for the usability experiments. Several tasks will be settled in order to test the usability of different UI components. Task 1 indicates the traditional approach to balancing utility and privacy, which is used as the baseline of the test. In task 2 we want to test the usability of the numerical indicator compared with the spreadsheet in task 1. Task 3 focuses on the difference between the trivial and PCC methods in selecting anonymization parameters. As the scatter plot is integrated with the auto PCC, we do not make a comparison between them. Task 4 as an entire approach will be compared with the other tasks in balancing utility and privacy. This structured approach allows for a comprehensive assessment of different anonymization strategy selection methods in the context of usability.

Task	**UI Components**			**Interactions**		
	Spreadsheet	**Indicators**	**Scatter Plot**	**Trivial**	**PCC**	**AutoPCC**
1	✓			✓		
2		✓		✓		
3		✓			✓	
4			✓			✓

Table 3.1: Bias of the comparison experiment. "Indicator" is the " Numerical indicators of Privacy and Utility"

Figure 3.6 indicates the sketch of the lo-fi prototype. In the low-fidelity prototype, we implemented the following functions:

1. **Data uploader**. Users can upload the data set that needs to be anonymized. In this function, we currently only support users to upload files in CSV format. The file upload function in other formats can be added in subsequent product versions. This low-fidelity prototype only implements basic functions.

2. **Original data presentation**. The data set is presented in a spreadsheet format. We also added some basic bar charts to show the overview of the data and distributions of individual columns as well (As shown in Figure 3.7(c) and (d)).

3. **Anonymization settings**. Since in the Mondrian algorithm, we need to set the quasi-identifiers (QIs), sensitive attribute(SA) of the data set, and the anonymization method (k-anonymity and its extension) as well as its parameters, the users have to select this information manually. The selectors we provide to the users include the **QIs and SA selector**, the selector of anonymization method (**anony-method selector**), and the **parameters selector**. There are two kinds of parameter selectors, namely **trivial parameter selector** and **PCC parameter selector**. The trivial parameters refer to k, l, t and the PCC parameter is the parameter p when the user applies the PCC method.

4. **Anonymized data presentation.** Here we provided the users three anonymization displays, including **anonymized data presentation**, **utility presentation**, and **privacy presentation**. We will reuse the spreadsheet and the bar charts to present the anonymized data. Numerical indicators and a scatter plot are utilized to demonstrate the utility loss and privacy loss after anonymization.

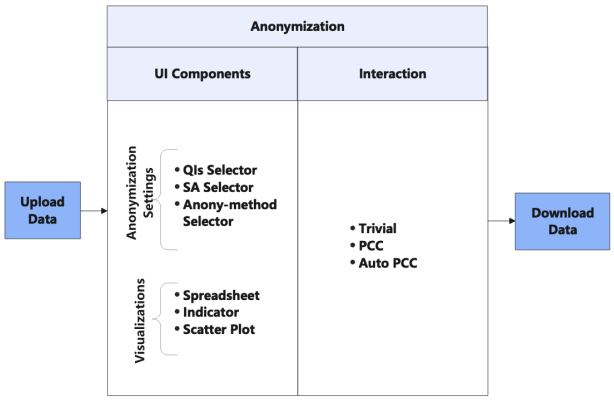


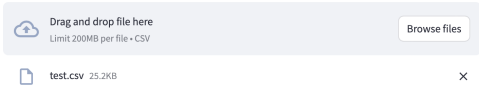
Figure 3.6: Sketch of the lo-fi UI

Interface Implementation

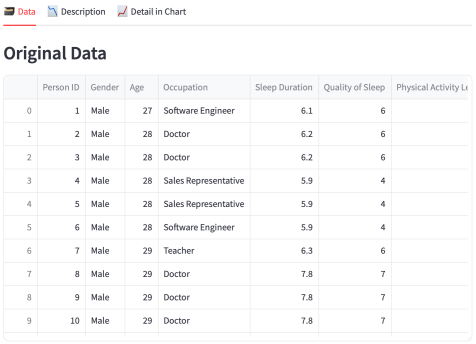
The whole widgets in the lo-fi prototype are provided by Streamlit. All the interactions of each component are provided by Streamlit automatically as well. In Figure 3.7, (b) (c) (d) indicate the tabs of the spreadsheet, which allows users to easily move between contents of the data. (b) is the original data, which shows the data in a table. (c) indicates the description of the data

in numerical and categorical data, which shows the statistical features of the data set. Through (d) the users can select the attributes they are interested in and visualize the selected columns of data in bar charts. (e) is the QIs and SA selector. The QIs selector is a multi-selector that allows users to select multiple items. The SA selector is a single selector, which allows users to select only one item. The setting of the number of QIs and SA is dependent on the Mondrian algorithm. (f) is the method and parameter selector. Since different methods have different parameters to modify, we set an interaction between the method selector and the parameter selector. For instance, if the user selects "k-anonymity" in the method selector, the parameter selector shows only one select slide of parameter k. (g) indicates the numerical indicator of utility loss and privacy loss of the anonymized data. The big number shows the loss and the small number indicates the change compared to the last choice of the user, which shows the loss increase with an up arrow in red and the loss decrease with a down arrow in green. (h) shows the scatter plot of the utility loss and privacy loss. Each point indicates one anonymization solution from the auto-PCC. The scatter plot shows all the anonymization solutions conducted by auto-PCC.

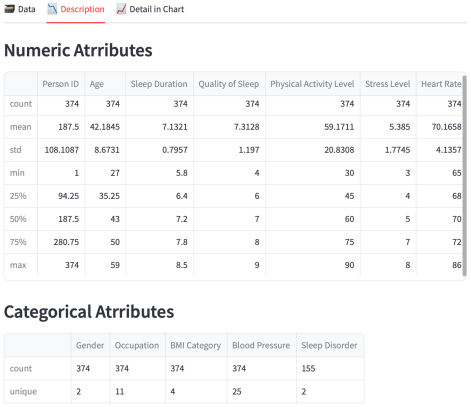
The layout of the prototype is up-to-down, which is embedded in different tasks of the first usability test. The original code is stored on github [lo-b] and the lo-fi prototype is deployed online [lo-a]. We will show the Results of the usability test using the lo-fi prototype in Section 4.2.1. Via the lo-fi prototype and the first usability test, we have found which UI components have positive effects on balancing utility and privacy trade-offs. Furthermore, after the first iteration of the prototype we will find out the issues regarding interaction and business logic.



(a) Data Uploader

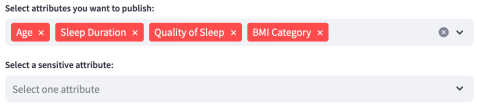


(b) Original Data



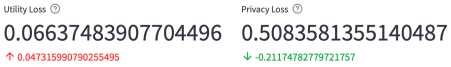
(c) Data Description

(d) Data in columns

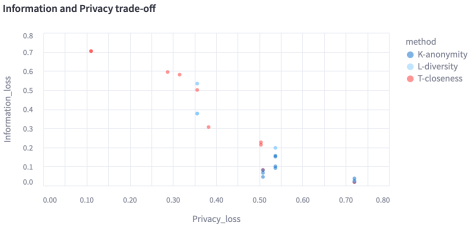


(e) QIs and SA Selector

(f) Method and Parameter Selector



(g) Numerical Indicator



(h) Scatter Plot

Figure 3.7: UI Components in Lo-fi Prototype

3.3.2 High Fidelity Prototype

In the hi-fi prototype, we mainly focus on the interactivity of the UI components and customization of the application. In the first usability test, we determined the usability of each component in the application and summarized the problems that occurred in the low-fidelity prototype through feedback from the test users. In the high-fidelity prototype, we made changes to address these issues.

Development Framework

Streamlit is mainly used for rapid prototyping development and rapid iteration of analysis tasks, it has limitations in interactivity and the application response time is relatively long. Therefore we use Dash [DAS] as the program framework for the high-fidelity prototype. Dash provides more flexibility and customization, giving developers more precise control over the appearance and behavior of their applications. Dash is built on Plotly [PLO] and Flask [FLA], which means we can easily integrate the functionality of Plotly charts and Flask, and for projects that require more advanced customization and extensions, Dash provides more freedom. With Dash, we can create applications with complex interactive elements, real-time updates, and responses. Furthermore, Dash applications can be embedded in other web applications, which provides more possibilities and application scenarios for this prototype in future work.

Interface Implementation

Based on the result of the first iteration usability test, we refined the tool's workflow and made adjustments to the UI components in the low-fidelity prototype. Figure 3.8 shows the work flow of the hi-fi prototype. Figure 3.9 Shows the interface of the hi-fi prototype. For instance running, online prototype is available [HIFb] and the github repository is [HIFa].

The high-fidelity prototype inherited most functions of the low-fidelity prototype. According to the feedback from the first usability test, we found that auto PCC can significantly improve the usability of the prototype. The feedback from the users indicates that the use of professional terminologies will make users confused (6 out of 10 test participants clearly expressed that It's hard to understand the professional terminology and it will disturb the decision-making), so we deleted the anony-method selector and the parameter selector Figure 3.7(f) in the high-fidelity prototype, which means the users are not required to manually select anonymization methods and parameters, but directly select the results of the auto-PCC from the scatterplot Figure 3.9(f). Second, we removed the data description Figure 3.7(c) and column bar chart Figure 3.7(d) functions of the spreadsheet because in the first usability test, only 1 out of 10 participants thought that the data description and bar charts were helpful for users to understand the data. Third, we display the numerical

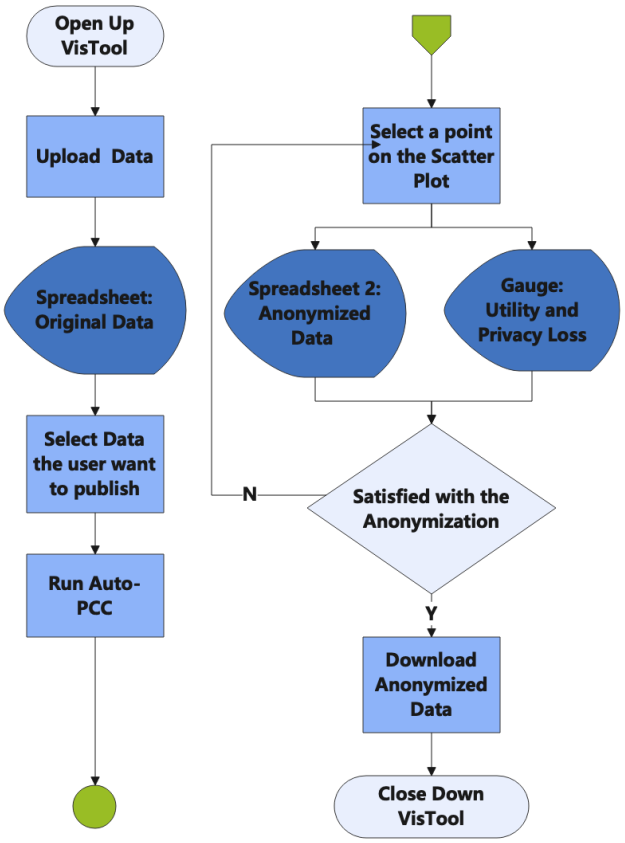
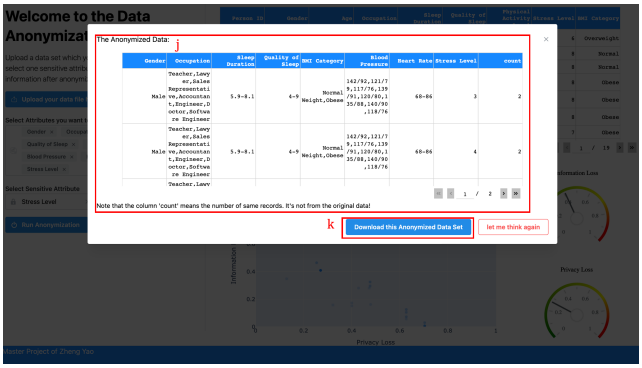


Figure 3.8: Workflow of the Hi-Fi Prototype

indicator Figure 3.7(g) in the form of two gauges Figure 3.9(g). In the gauge, the low loss is represented in green, and the high loss is represented in red. The gauge intuitively shows that the value domain of the loss is [0,1], which helps users understand utility loss and privacy loss. (We added a help tip in the low-fidelity prototype, but the users rarely clicked it. In addition, we replaced ”utility” with ”information” because, in the usability test, 3 users had a poor understanding of ”utility” Confusion arose. Fourth, the dash framework greatly improves the response speed of the prototype, which is also an obvious problem in the low-fidelity prototype. Finally, we added interaction to the scatter plot, allowing users to directly click the point on the scatter plot to select an anonymization strategy. In the low-fidelity prototype, the user needs to manually adjust the anonymization method and parameters based on the information provided by the scatter plot while the high-fidelity prototype simplifies this process. After the user clicks a point, the download interface pops up in the form of a floating window, which provides the user a spreadsheet with anonymized data Figure 3.9(j). Then the user can download this anonymized data set with one click on the download button Figure 3.9(k).

(a) Main Interface



(b) Download Interface

Figure 3.9: High Fidelity Prototype. (a)Introduction. (b)Data uploader. (c)QIs and SA selector. (d)button to run auto-PCC. (e)Original data spreadsheet. (f)Scatter plot. (g)Two gauges of utility and privacy loss. (h)Process status of auto-PCC. (i)Footer. (j)anonymized data spreadsheet. (k)Data downloader.

3.3.3 Optimized Hi-Fi Prototype

Figure 3.10: Modified Interface. (a)Attributes Selector. (b)Notification.(c)Title of the Scatter Plot

In response to the issues identified in the second iteration of the usability test, several enhancements were implemented in the prototype. Firstly, to address user confusion about the Sensitive Attribute (SA), we set it to a default value, specifically the last value of the selected attributes. Considering the results in **??**, indicating that the prototype's performance is generally robust across different SA values, this adjustment aims to streamline the user experience. The SA selector is now hidden within the "More Setting" section (Figure 3.10 (a)), presented as a pop-up drawer window Figure 3.11, allowing users with an interest in understanding SA to make adjustments. Future iterations may introduce additional advanced options in the drawer, such as choosing different anonymization parameters and adjusting the iteration count for PCC.

Secondly, a notification feature was introduced (Figure 3.10 (b)). Users receive notifications in the bottom-left corner of the page when they click the "Anonymization" button, providing real-time feedback on the anonymization process Figure 3.13. Upon completion, a notification informs users that the anonymization process has finished. Lastly, modifications were made to the Scatter plot's title (Figure 3.10 (c)), removing the reference to clicking on points. Instead, interactive elements were added to the Scatter plot. When users hover over data points, a tooltip appears, guiding them to click on the points to download anonymized data (Figure 3.12).

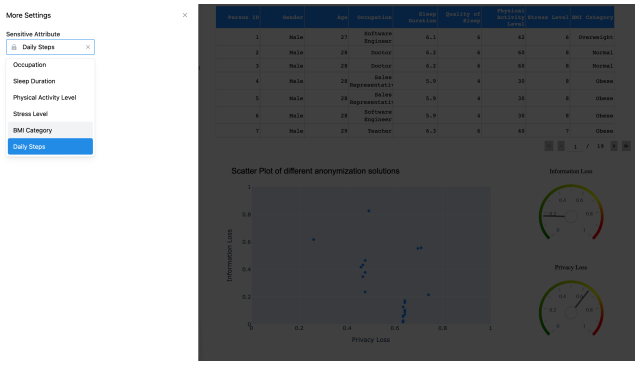


Figure 3.11: Drawer

As we move forward, the implemented solutions pave the way for the subsequent evaluation and testing phases, where the efficacy and performance of our system will be rigorously examined.

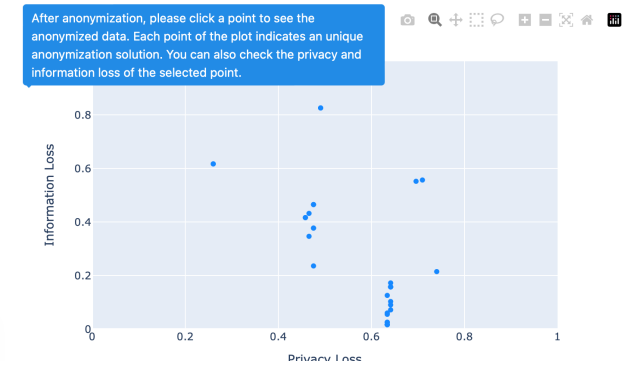


Figure 3.12: Scatter Plot (Modified)

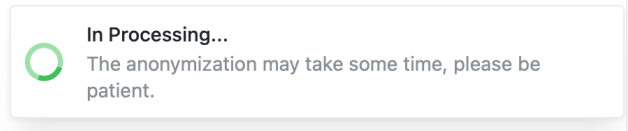


Figure 3.13: Notification: In Processing

4 Evaluation

Our evaluation comprises two core aspects. Firstly, the case study delves into the prototype's functionality, emphasizing the effectiveness of PCC method. Secondly, the usability test explores the UI's evolution through three design iterations, showcasing substantial improvements. Our evaluation endeavors to present a holistic view of the prototype's strengths, from its foundational methodology to its functional prowess and enhanced usability. Through this multifaceted evaluation, we aim to underscore the effectiveness and reliability of our data anonymization solution.

4.1 Case Study

The case study initiates with the application of various datasets to our prototype, rigorously testing the functionalities of auto PCC and scatter plot. Subsequent scrutiny aims to assess the effectiveness of the developed prototype, specifically focusing on its impact on achieving a balance between utility and privacy. Through these objectives, the case study seeks to offer a comprehensive evaluation of the prototype's practical implications.

The primary aim of this section is to substantiate the positive impact of the prototype in achieving a favorable trade-off between utility and privacy. Our hypothesis posits that the prototype, when applied to diverse data sets, will exhibit a reduced concentration of data points near the origin in scatter plots (as illustrated in Figure 1.1, a=5). This reduction signifies a successful mitigation of privacy concerns while preserving data utility. By analyzing the dispersion of data points, we seek to demonstrate the prototype's ability to navigate the trade-off effectively.

4.1.1 Datasets

According to Section 2.4.1, we selected four data sets from Kaggle[kag]:

1. **Sleep Health and Life style**[Tha]. This Data set comprises 400 rows and 13 columns, covering a wide range of variables related to sleep and daily habits. It includes details such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders. Most of the attributes are numerical and there is no obvious correlation among the attributes.

2. **Flight Price Prediction**[Bat]. This dataset contains information about flight booking options from the website Easemytrip for flight travel be-

tween India's top 6 metro cities. There are 300261 data points and 11 features. Most features are categorical.

3. **Complete Historical Cryptocurrency Financial Data**[Pmo].Consolidated financial information for the top 10 cryptocurrencies by market cap. Pulled from CoinMarketCap.com. Attributes include Currency name (e.g. bitcoin), Date, Open time, Highest price, Lowest price, Close time. Most of the variables are numerical.

4. **Pokémon Legendary Data**[Awa]. This data set has a series of numerical fighter stats: attack, defense, speed, and so on – as well as a categorization of Pokemon type (bug, dark, dragon, etc.).

The dataset selection aligns with the methodology's requirements (Section 2.4.1), showcasing diversity, sufficient quantity and quality, and attribute relevance. The datasets chosen cover a range of topics. Their respective scales provides a solid foundation for comprehensive analysis.

4.1.2 Test Design

In the context of data anonymization, our selected datasets exhibit a diverse range of features that are crucial for analyzing the effectiveness of our prototype. We focus on several key characteristics:

1. **Data Type**. The datasets encompass both numerical and categorical data, allowing us to assess the prototype's performance in handling different data types.

2. **Correlation Between Attributes**. Variability in the correlation between data attributes is observed. For instance, data set 3 [Pmo] displays strong correlations among certain data attributes (Highest price, lowest price, etc.), while others lack significant correlations. This analysis helps evaluate the prototype's robustness in managing correlated data.

3. **Quasi-identifiers (QIs) Quantity**. Assessing the number of quasi-identifiers provides insights into how the prototype performs at different scales. Understanding its effectiveness across varying quantities of quasi-identifiers is essential for scalability considerations.

4. **Sensitive Attribute Value Range**. The datasets present varying sizes of value ranges for SA. Some SAs may have a binary value, while others may exhibit a larger range. Investigating the impact of the SA value range on the prototype's effectiveness is an aspect of our analysis.

We set up seven preliminary tests (in Table 4.1) to see if these features affect the result of the visualization. We believe that these variety of data features can simulate most scenarios in our daily usage.

Test Nr.	QI-type	QI-number	QI-correlation	SA-value domain
1	N	5	low	large
2	N	2	low	large
3	N	10	low	large
4	C	5	low	large
5	N	5	high	large
6	N and C	5	low	large
7	N	5	low	small

Table 4.1: Design of the tests. "N" and "C" indicate "numerical" and "categorical". Each column represents the features of data utilized in the tests.

Because of the Mondrian algorithm, the SA would not be changed in the anonymization process. Since the value domain will affect the value of parameter l_{max} in the anonymization process, we just focus on the value domain of the SA and observe the test result under a small or large value domain. Tests 1,2,3 focus on the effect of the number of quasi-identifiers. We set the number values 2, 5, and 10 to simulate different cases. We set Test 1 as a baseline, which will be compared with the other tests. For instance, Tests 4, and 6 focus on the data type, and Test 5 focuses on data correlation. Test 7 focuses on the attribute value domain of the SA. We determine the data and attribute selection for each test in Table 4.2:

Test Nr.	Dataset	QIs	SA
T1	D1	Age, Sleep Duration, Physical Activity Level, Heart Rate, Daily Steps	Stress level
T2	D1	Age, Sleep Duration	Stress level
T3	D4	Attack, Defense, Height, Hp, Percentage male, Sp attack, Sp defense, Speed, Weight, Generation	Type
T4	D2	Flight, Source city, Departure time, Stops, Arrival time	Destination city
T5	D3	Open, High, Low, Close, Volume	Currency
T6	D1	Age, Sleep Duration, Physical Activity Level, BMI Category, Sleep Disorder	Stress level
T7	D1	Age, Sleep Duration, Physical Activity Level, Heart Rate, Daily Steps	Gender

Table 4.2: Data and Attributes selection of the Tests. D1: Sleep health and life style[Tha]. D2: Flight price prediction[Bat]. D3: Complete Historical Cryptocurrency Financial Data[Pmo] D4: Pokemon Legendary Data[Awa]. T1 is the baseline test. T2 is the test with small QIs. T3 is with large QIs. T4 is the test with categorical QIs. T5 is the test with high QIs-correlations. T6 is the test with numerical and categorical QIs. T7 is the test with small SA value domain.

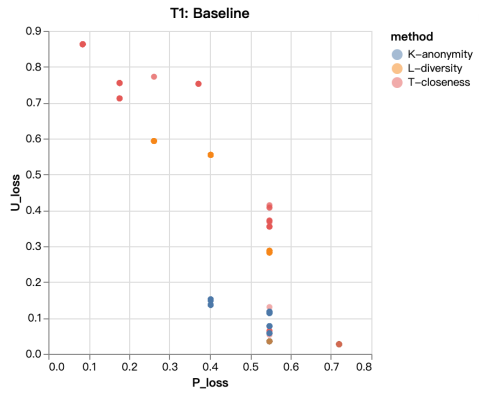
4.1.3 Results and Analysis

In this section, we incorporated data from these 7 test plans into our prototype, generating scatter plots that encapsulate the outcomes of 7 tasks, each comprising 100 data points Figure 4.1). These plots illustrate the distribution of 100 data points concerning the trade-off between privacy and utility losses. Each data point represents a unique anonymization strategy, incorporating different anonymization methods and parameters, with distinct colors denoting various anonymization methods. The x-axis and y-axis represent privacy loss and utility loss, respectively, ranging from 0 to 1, indicating increasing degrees of loss. Proximity to the origin signifies that the represented privacy strategy minimizes both utility and privacy losses, reflecting a better balance. The negative correlation between privacy and utility losses shown on the plots aligns with real-world expectations. While points are evenly distributed across the entire domain, there is clustering along specific values on the x-axis, potentially influenced by the chosen measurement method: variations in different information entropy may result in the same JS divergence. Consequently, although some anonymized datasets exhibit distinct privacy disclosures, they share similar privacy loss values in measurements. The overarching observation reveals a commendable distribution pattern across all scatter plots. Notably, the points uniformly populate the utility loss (y-axis), forming discernible "vertical stripes" when clustered on the privacy loss (x-axis).

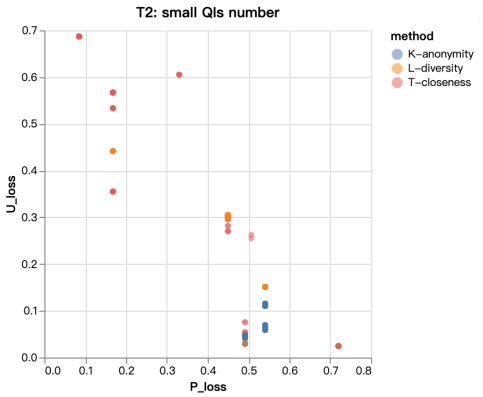
Within each stripe, points share identical privacy loss values while exhibiting diverse utility loss values. This structure allows users to easily identify "local optimal points" at the bottom of each stripe. Given the negative correlation between privacy loss and utility loss, selecting a local optimal point involves a trade-off between the two factors, enabling users to align with their specific privacy-utility requirements.

Tests 1-6 exhibit consistent scatter plot characteristics, showcasing even distribution along the utility loss axis and distinctive clustering on the privacy loss axis. However, Tests 7 and 8 display distinct patterns. Figure 4.1(g) and (h) in particular exhibit outliers along the axis, introducing ambiguity. Despite an even distribution within a limited range, points cluster and overlap, suggesting minimal changes in anonymization outcomes across different methods and parameters.

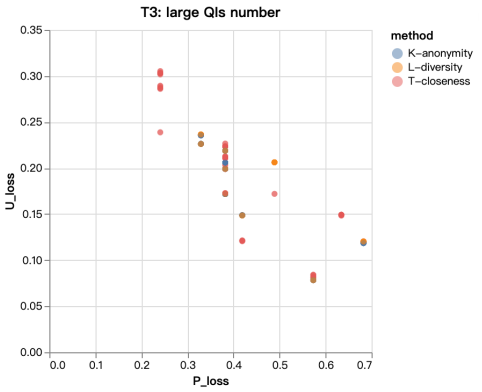
However, it is crucial to emphasize the limitations inherent in our results. The reliance on scatter plots, while insightful, may not capture all nuances of the data distribution. The visual analysis, though valuable, is limited in its ability to uncover subtleties that might be discernible only through more sophisticated statistical methods. Additionally, the interpretation of "optimal points" is subjective and context-dependent. The user's definition of an optimal trade-off between privacy and utility may vary, and our approach might not cater to all possible user preferences. In order to alleviate these limitations, different UI components need to be applied to the tool to help users from different perspectives (such as gauge and spreadsheet introduced in Section 3.3).



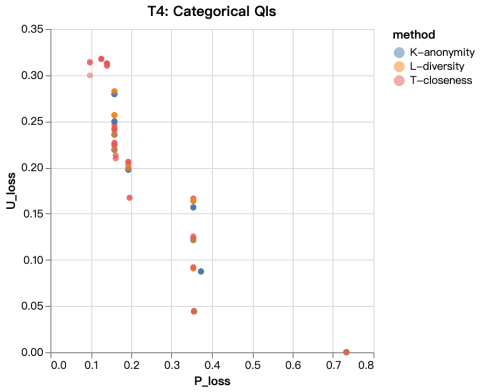
(a) T1



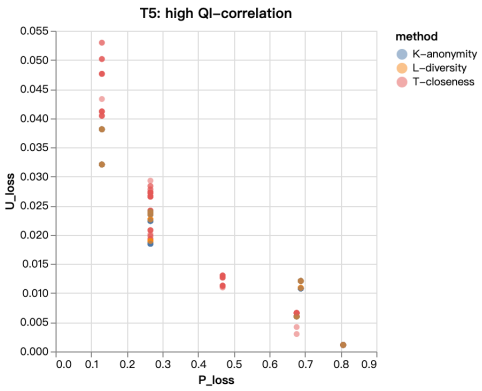
(b) T2



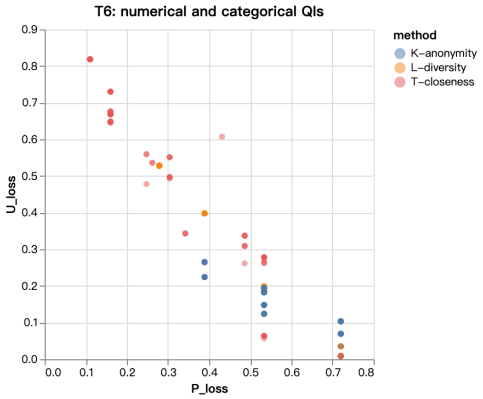
(c) T3



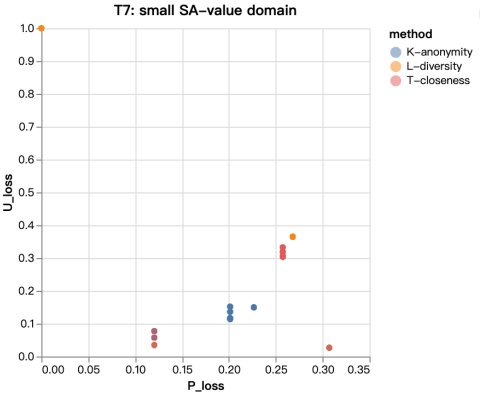
(d) T4



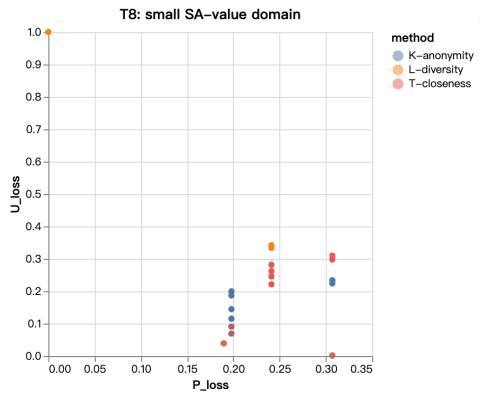
(e) T5



(f) T6



(g) T7



(h) T8

Figure 4.1: Privacy-utility Trade-off Scatter Plot of 8 Tests

The outliers observed in Tests 7 and 8 introduce an element of uncertainty, and their impact on the overall conclusions warrants cautious consideration. It is essential to acknowledge that outliers can significantly influence the interpretation of results and may be indicative of underlying issues or anomalies in the anonymization process.

Our innovative features, Auto PCC and scatter plot, showcased notable integration of utility and privacy in the case study results. In the majority of instances, utility and privacy were effectively visualized in the same interface, providing users with actionable guidance in achieving a balanced trade-off. This directly addresses our research question, RQ1: "How should data controllers consider privacy and utility in an integrated framework for privacy-preserving data publishing?" Our thorough analysis of the scatter plots reveals nuanced trade-offs, empowering users to select optimal points tailored to their preferences. While our approach demonstrates robustness across diverse test scenarios, the findings should be interpreted within the study's specific conditions. Caution is advised in extending conclusions, and further research with varied datasets and refined methodologies is essential to enhance the generalizability of our approach.

4.1.4 Further Discussion

In our further exploration, we sought to compare the effectiveness of different anonymization methods, namely k-anonymity, l-diversity, and t-closeness. Applying quantitative metrics from the methodology (Section 3.2.3) to assess the trade-off effects, we aimed to understand the specific performance of each method at varying privacy levels. The results demonstrated significant variations in the performance of these anonymization methods.

In the seven tests conducted during the case study, we computed the effectiveness (Section 3.2.3) for all 100 data points, illustrating the results through bar charts as shown in Figure 4.2. The charts display the effectiveness of the 100 samples under various anonymization strategies. The x-axis represents the parameter p values for PCC, ranging from 0 to 1, indicating increasingly stringent privacy requirements. Different anonymization methods are distinguished by color. The y-axis represents effectiveness. However, it's crucial to note that effectiveness is not linear (As shown in Figure 3.5). This implies that absolute effectiveness values cannot be directly compared, allowing for relative comparisons only. Figure 4.2 presents bar charts illustrating the effectiveness under different anonymization methods, accompanied by changes in the p-value. The experiments conducted (Test 1, 2, 6, and 7) were derived from the same dataset ([Tha]). Surprisingly, each anonymization method exhibited distinct behavior under different quasi-identifiers (QIs) and sensitive attribute (SA) scenarios, even when applied to the same dataset.

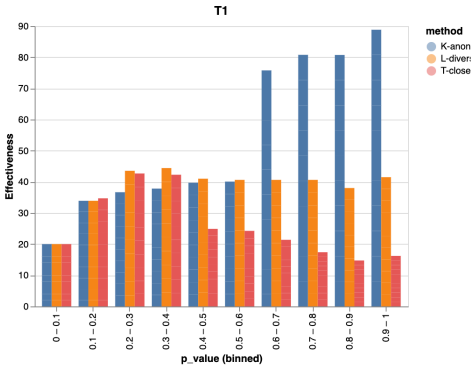
For instance, in Test 1 and Test 2, where the SA remains consistent, the effectiveness of k-anonymity stabilizes within $p \in [0, 0.6]$, showing minor differences among the three methods. However, when p $>$ 0.6, the k-anonymity

method significantly outperforms l-diversity and t-closeness, reaching a peak in $p \in [0.9, 1]$. In contrast, Test 2 reveals k-anonymity as the least effective method.

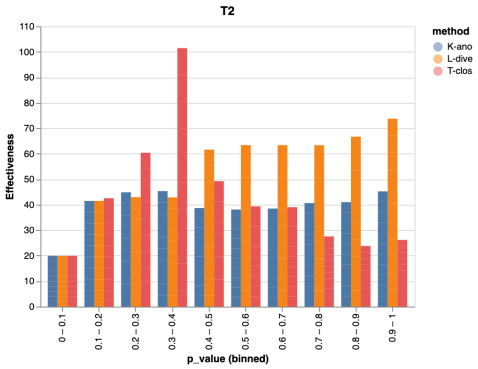
Test 6 mirrors the results of Test 1, indicating the sensitivity of anonymization methods to specific QIs and SA-choosing cases. Test 7 showcases high effectiveness in the anonymization process under k-anonymity and t-closeness in $p \in [0.1, 0.4]$, while l-diversity performs well in $p \in [0.1, 0.2]$ but quickly diminishes.

Importantly, there is no overarching principle guiding data controllers on the optimal anonymization method for specific scenarios. Each method exhibits varying behavior, significantly impacting effectiveness at specific privacy levels (e.g., k-anonymity in Test 1 with $p \in [0.7, 1]$ and t-closeness in Test 2 with $p \in [0.3, 0.4]$), or proving less effective in most privacy requirements (e.g., l-diversity in Test 7 with $p \in [0.5, 1]$). The concentration of high effectiveness in a narrow range in Test 7 indicates the limiting effect of our anonymization process on the data.

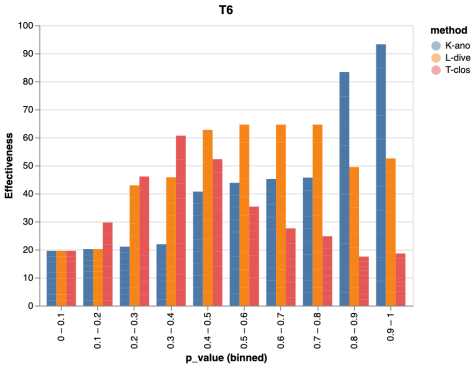
In summary, the absence of a universal answer to the optimal anonymization method underscores the contextual nature of effectiveness, necessitating a tailored analysis for each scenario. This highlights the limitations of the traditional heuristic approach, which may prove ineffective when handling diverse datasets. On the contrary, the diverse outcomes from our experiments validate the adaptability and robustness of the proposed tool across different datasets.



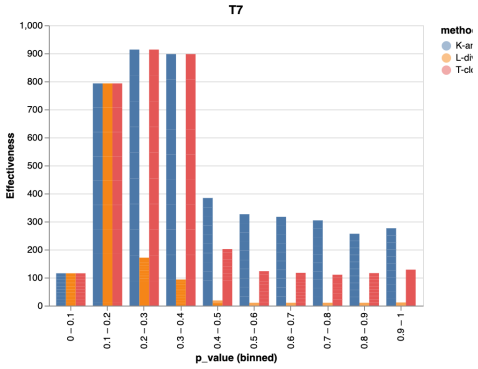
(a) Test 1



(b) Test 2



(c) Test 6



(d) Test 7

Figure 4.2: Bar charts: Effectiveness distribution on various anonymization methods

4.2 Usability Test

Motivated by the imperative to address RQ2 — determining the UI components that enhance the usability of our data anonymization tool for data controllers seeking the optimal utility-privacy trade-off, we embarked on a rigorous usability testing process. Recognizing the significance of a user-centric approach, we initiated a comprehensive evaluation of our tool's usability prior to its public release. The initial phase involved a low-fidelity version of the tool, assessing users' navigation ease through the anonymization process. This approach allowed us to pinpoint the UI components crucial for enhancing overall usability and aiding data controllers in their quest for the most effective utility-privacy balance. The motivation behind this undertaking lies in the desire to refine and optimize our tool, ensuring it seamlessly integrates with other UI componrnts to form a cohesive and user-friendly data anonymization solution. Through this usability testing initiative, we aspire to uncover valuable insights, rectify design issues, and ultimately deliver an enhanced product to our user community.

Subsequently, we transition to the interactive test phase, optimizing the interaction design of the UI within a high-fidelity version of the tool. This phase is dedicated to identifying user challenges during operation, with a specific focus on interaction methods and information retrieval processes. Finally, based on the insights gathered from these two phases, we refine the high-fidelity prototype to address any issues identified during the second iteration of testing. A third iteration of testing is then conducted to validate the effectiveness of the adjustments made. Figure 4.3 indicates the Phases of the usability test.

In summary, our usability testing initiative is strategically designed to assess the effectiveness of various UI components within the data anonymization tool, with the overarching goal of seamlessly integrating these components into a cohesive and user-friendly solution. By meticulously evaluating the usability of individual elements and their collective impact, we aim to develop a tool that not only functions optimally but also aligns with user expectations. This comprehensive approach ensures that the data anonymization tool goes beyond mere functionality, addressing the nuanced needs of data controllers as they navigate the complex landscape of achieving a balanced utility-privacy trade-off. The insights garnered from this usability testing process serve as a pivotal guide in refining and optimizing the tool, ultimately contributing to the successful resolution of RQ2 and the development of a robust, user-centric data anonymization solution.

Since the target users of our tool are data controllers, which has no limitation on a specific type of people, we just make sure that the distribution of gender, age, and related data experience in the user sample is roughly even, so that the test results will not be greatly disturbed. We give the scale of data experience as follows:

- **Data newbies**, who have no experience in working with data. They

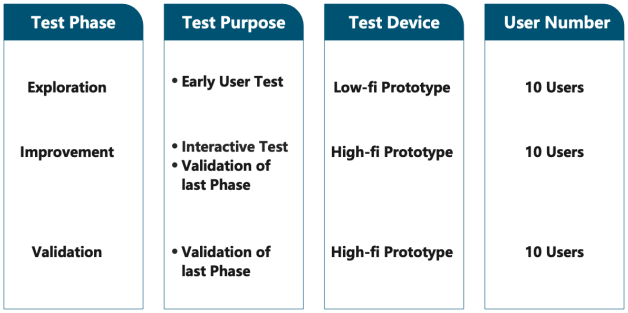


Figure 4.3: Phases of the usability test

also have less computer skills. For example musicians, and housewives.

- **Data beginners**, with little experience working with data. They also have some experience with certain systems or pieces of software, such as sociology students.
- **Data experts**, professionals working with data who have a lot of expertise. They are skilled in computers. For example, students majored in data science and data analysis.

The prototype is deployed online and all phases of the usability test will take place remotely. We use videoconferencing software e.g. Zoom or Webex to record user interactions, facial expressions, and audio comments in real-time. Figure 4.4 indicates the process of the test. The time cost of each user is 30 to 35 minutes. At the beginning of the test, the moderator briefly introduces the purpose and process of the test to the users to put the user in a more relaxed state. Afterward, the user needs to fill in a simple questionnaire about the personal information, the user's experience, and habits in using related products. Then the test begins. The moderator observes the user's expression and body movements while the user completes the task and asks some simple verbal questions after completing the task. After each task session, the user fills out a post-task questionnaire. In this study, we use the ASQ (After Scenario QuestionnaireSection 2.3.2) to investigate the usability of each single task session. Then we proceed to test the next task. In the second and third iterations, after all task sessions are completed, the user needs to fill in a questionnaire about overall satisfaction. We use SUS (System Usability Scale Section 2.3.1) in this research.

4.2.1 Exploration

In this section, We introduce the process of the first test iteration. We are aiming to find the optimal combination of UI components to improve the usability of the prototype. The test is designed as a series of task scenarios. In this phase, all the basic functions of the tool should be tested. The following are the tasks of the tool:

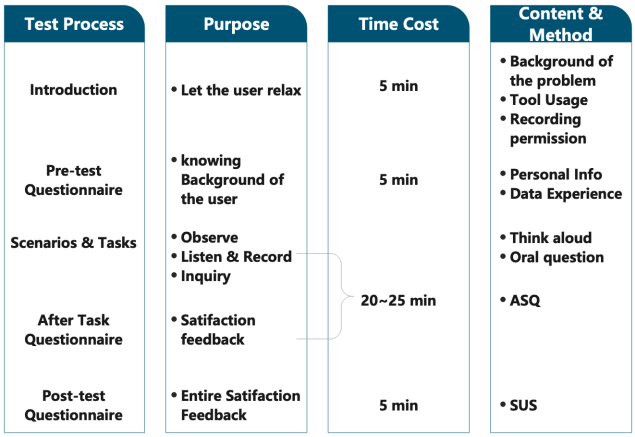


Figure 4.4: Process of the usability test

- **Task 1: upload a data set.** First, you should upload a data set. You may upload the given test data set (or your own csv data file).

- **Task 2: Select the attributes you want to publish.** You need to select some attributes of the data that you wanna show to the public. The anonymization process will mask these attributes to protect privacy. Among these attributes, there is a sensitive attribute that will not be masked. Your task in this step is: 1. Select some attributes of the data that you want to publish. 2. Select a sensitive attribute.

- **Task 3: Select your anonymization parameters-1** Now, you need to select a method for anonymization. We provide k-anonymity, l-diversity, and t-closeness. To get a good anonymization result, you need to: 1. Select an anonymization method. 2. Modify the parameters of the selected method. Check the anonymized data in the spreadsheet. 3. Redo steps 1 and 2 to get a better balance between privacy and information loss.

- **Task 4: Select your anonymization parameters-2** We provide you here an indicator of the utility- and privacy-loss. With the help of this indicator, maybe you have a new idea to modify your anonymization parameters. If so, please re-modify the anonymization method and parameters. You can also insist on your choice from the last task.

- **Task 5: Select your anonymization parameters-3** In this task, we integrate the parameter k, l, and t in one parameter p, which present low privacy to high privacy from 0 to 1. Modify parameter p to get an optimal balance of privacy and utility you think.

- **Task 6: Select your anonymization parameters-4** In this task, we run the parameter p from 0 to 1 automatically. The following visualization shows information loss and privacy loss in one scatter plot. With

the help of this scatter plot, please select an optimal balance of privacy and information loss you think.

Figure 4.5 shows the structure of all the tasks and their purposes. Tasks are separated into two categories: Basic UI and anony-UI. Tasks 1 and 2 aim to test the usability of two necessary UI components, including a data uploader and a selector of sensitive attributes and quasi-identifiers. These two components are the basic components in the workflow of anonymization, which are simple in logic. These two tasks play two roles in the whole test:

1. **Let the users get familiar with the test process.** With the help of these two tasks, the users can calm down quickly and get familiar with the test process.

2. **Provide users a frame of reference for comparison**. Since these tasks are the simplest of all 6, the ER and ASQ scores may be higher than the scores of the other tasks. Because the users have different grading scales. With the help of these two tasks, we will know the usability scale of the users. The users can also use the grade of these two tasks as a reference for grading.

The purpose of the other 4 tasks is to compare the usability of different UI components to help us find which components can improve the usability of such an anonymization tool. We propose here 4 UI components for anonymization, including a spreadsheet, a numerical indicator of privacy and utility, PCC(introduced in the section "Implementation"), and auto PCC with a scatter plot. Biases of the comparison are shown in Table 3.1. In the first iteration

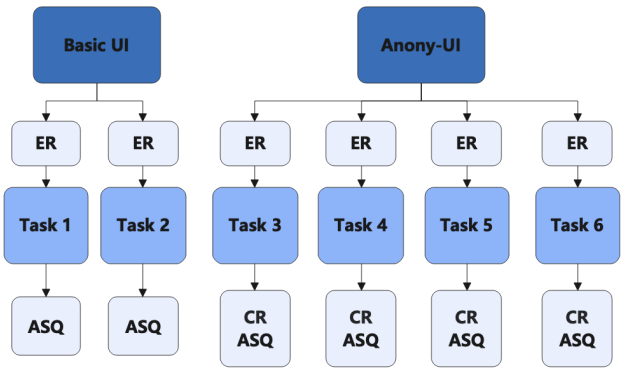


Figure 4.5: Structure of Tasks in Exploration. ER = expectation rate, ASQ = after scenarios questionnaire, CR = confidence rate.

of the usability test, we need 3 types of questionnaires: Expectation Rate(ER), After-Task Questionnaire(ASQ), and Confidence Rate(CR). The expectation rate draws a rate between how easy or difficult it is to act on the task before this task began. The question of ER is: " How do you expect the ease of this

task?". The score of ER will be compared with the score of the first subquestion of the ASQ because both these two questions are aimed at the difficulty of the task.

Test Participants

The first iteration of our usability testing comprised 10 participants selected through diverse channels, including households, online platforms, and personal connections. To ensure a broad spectrum of participants, we actively sought individuals from various age groups, genders, educational backgrounds, and with varying levels of experience in data analysis. By gathering participants through different avenues and pre-screening their basic demographic information, we aimed to prevent any concentration within a specific demographic. This approach was consistently applied across all three iterations of usability testing, ensuring a representative and varied participant pool throughout the evaluation process. Figure 4.6 shows the distribution considering these aspects. Since the data controller is not a group of people with specific attributes (such as age, occupation, gender, etc.), we will only investigate some basic personal information when selecting test participants and ensure that these attributes are evenly distributed among the test participants to ensure these basic properties do not potentially affect experimental results. The age of the participants is concentrated around 30 years old, with the youngest participant being 20 years old and the oldest being 65 years old. These participants were relatively evenly divided by gender and had varying data processing experience. Of the ten participants, eight had a bachelor's degree or above, and two had only a high school degree.

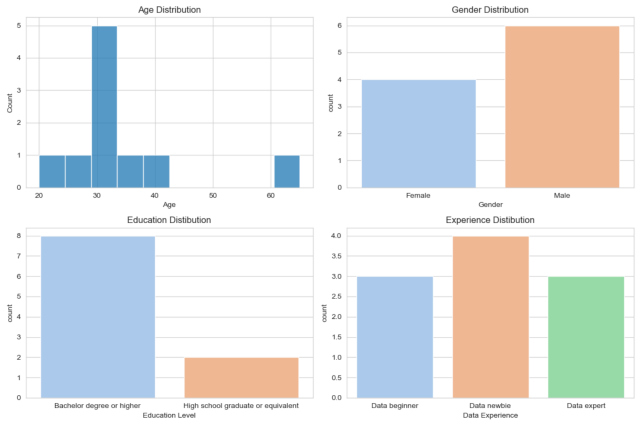


Figure 4.6: Participants distribution in the first usability test.

Expectation Rate

The steps of ER are like this: At the beginning, we introduce the UI components involved in the task to the user. Then the user gives the expectation

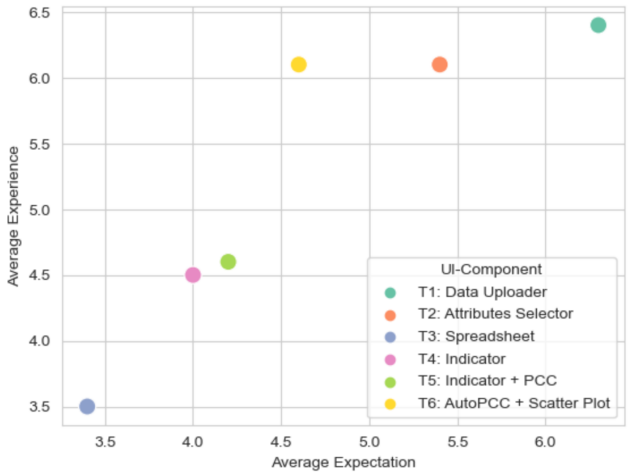


Figure 4.7: Expectation measurement. "Indicator" means "Numerical Indicators"

score to the components. After using the components, the user grade the experience score. Based on the results of the Expectation Rate (ER) shown in Figure 4.7, all tasks fall within the "Do not touch it" quadrant, as both the experience and expectation ratings exceed 3.5 points (refer to Figure 2.6). This suggests that overall, the UI components perform well in their respective tasks. However, a detailed comparison reveals interesting insights. Task 1 stands out as the top performer, with an average expectation and experience rating 6.5 of 7, aligning with expectations given that uploading a file is the simplest task of all. Task 2 follows closely, performing as expected. Tasks on the Anony-UI exhibit variations against each other, with Task 3 (representing the spreadsheet) performing the least favorably. The user expectations and experience hover around 3.75, close to the borderline of the "Don't touch it" zone. Task 4 shows improvement over Task 3, and Task 5 exhibits a performance similar to Task 4. This indicates that the numerical indicator proves more usable than the spreadsheet, with manual PCC showing only slight improvement over traditional anonymous parameters. Surprisingly, Task 6, representing the Anony-UI component of auto PCC and scatter plot, outshines all others. Its performance is comparable to Task 2, suggesting that auto PCC and scatter plot effectively simplify a complex problem, akin to Task 2's simplicity (selecting QIs and SA parameters from a dataset). Moreover, Task 6's user experience surpasses expectations by 1.5 points, indicating that this UI component's actual impact exceeds user expectations. This makes it a more competitive solution for data anonymization and the privacy-utility trade-off. From the result of ER, the broad conclusion is that all components, even the traditional spreadsheet method, provide acceptable user experiences.

Confidence Rate

Overall, the users' confidence in using Anony-UI components shows an increasing trend from Task 3 to Task 6, which fits our expectations about auto PCC and scatter plots. However, this trend is not monotonically increasing.

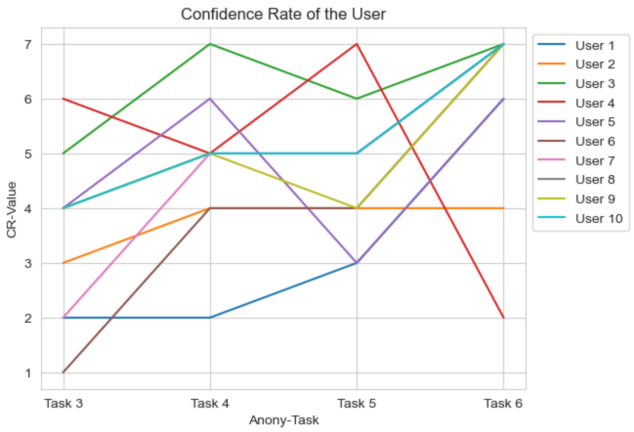


Figure 4.8: Confidence rate of users on anony-UI

As shown in Figure 4.8, the line of user 1 increases as the test goes from task 3 to task 6, which is of course an ideal experimental result. The performance of other users is different. We notice that the confidence of all users (except user 4) increased from 1 to 2 CR values from task 3 to task 4, which indicates that the numerical indicator has a great positive influence on the user's choice. However, comparing Task 4 and Task 5, we found that the confidence not only did not increase but remained the same or even decreased. This shows that the manual PCC does not have a positive impact on the choice of balancing utility and privacy from the perspective of the users, and even causes interference. In the experiment, users believed that indicating parameters k, l, and t into one parameter p did not simplify the parameter selection. Users (except data experts) said they do not understand the principles of k-anonymity, l-diversity, and t-closeness, no matter if it is a single parameter p or several parameters k, l, or t, it will confuse them. Therefore, manual PCC does not bring improvement in user confidence. When we switched from manual PCC to auto PCC and scatter plot, the confidence levels of 5 participants improved significantly. However, the confidence of users 2, 3, and 5 only returned to the same level as the numerical indicators. This indicates that for these users, auto PCC and scatter plots do not lead to better confidence than the numerical indicator. In the experiment, these participants stated that scatter plots did not interact with the parameter selector, which confused the users. These users said they have no doubt that the scatter plot shows all anonymization possibilities. However, they said these possibilities sometimes do not have obvious differences, which makes it difficult for them to make a decision. The users also felt that making choices based on scatter plots did not make the

results more reliable. This experimental result suggests that we should solve this interaction problem in the second iteration. User 4 was confused by the tasks throughout the whole test process. They struggled throughout the testing process with the principles of the anonymization methods(k-anonymity, l-diversity, and t-closeness). They said they didn't understand the methods, so they had no way of making the right decision. This also reminds us that even mentioning the name of an anonymization method will make users feel uncomfortable and confused. This issue should also be solved in the second iteration.

Effectiveness

During the test, we recorded the user's choices for anonymization and its parameters and calculated the effectiveness based on the mathematical model we proposed in Section 3.2.3. Figure 4.9 shows the average effectiveness value of all participants under tasks 3 to 6. Because the attributes of the data set that the user selected are different, the effectiveness value obtained by the user will also be greatly different (the value domain of effectiveness is $[0, +\infty)$). Overall, the effectiveness continues to increase from Task 3 to 6, and the effectiveness in Task 6 is much greater than the value under other tasks, which indicates that auto PCC and scatter plot will significantly help users obtain better solutions than the traditional approach. It is worth noting that the actual effectiveness and user's confidence are very different in different tasks (the user's confidence in Figure 4.8 is roughly the same at task 6 and task 4). Even if the user has great doubts about the choice based on the scatter plot, the user made a good choice. In the experiment, users' doubts about their choices since they thought they are lack of understanding of the principles of the anonymization process. The actual effect of each anonymization method shown in the scatter plot was very different from what the users expected. For example, user 3 claimed in the experiment that they thought t-closeness should give more reliable results (As t-closeness is developed from k-anonymity), but in fact, the best point selected based on the scatter plot was the k-anonymity method. User 7 believes that the anonymization parameter p should achieve a good balance of utility and privacy at 0.5, but the point they selected on the scatter plot is p=1. These users' presets of the results will affect users' judgment when applying scatter plots, so we will dilute this "stereotype" for anonymization methods in the second iteration.

After Scenario Questionnaire

Figure 4.10 shows the average ASQ values of 10 participants in each task. There is no doubt that tasks 1 and 2 have the highest ASQ values, close to 7 points and 6.5 points respectively. Tasks 3, 4, and 5 are basically in line with our expectations that the overall score from Task 3 to Task 5 is higher and higher. However, the ASQ values of task 6 behave completely differently.

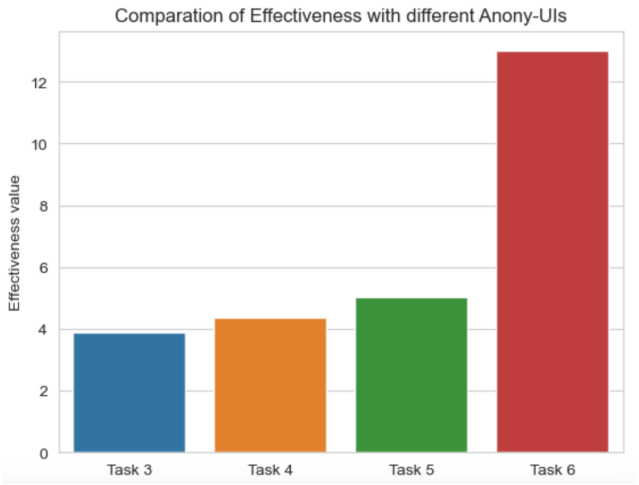


Figure 4.9: Effectiveness comparation

First of all, the Effectiveness value of task 6 is very high, reaching about 6.5 points. This shows that users give high ratings to the Effectiveness of auto PCC and scatter plot to complete their tasks. Secondly, the efficiency value of task 6 is also in line with our expectations. Compared with other Anony-UI components, auto PCC and the scatter plot significantly improve the time consumption of completing tasks. Finally, the satisfaction value of task 6 did not receive as high a score as effectiveness and efficiency. And compared to other UI components, user satisfaction is also at a low level. An interesting phenomenon is that there is a contradiction between CR values(Figure 4.8) and user satisfaction results in the ASQ by Task 5 and 6. It is hypothesized that inappropriate interactions may lead to a decrease in user satisfaction, even though the components themselves can offer high confidence levels. During the experiment, users generally claimed confusion about the interaction logic of this component. Users experience frustrations during the interaction process, such as difficulties in understanding and navigating the interface. These challenges might not necessarily reflect on the functionality or accuracy of the components but can significantly impact the overall user experience. We will address this issue in the second iteration.

Issues occurred in the Test

Through the observation recorded during the test process, we extracted the users' questions and subjective feelings, and summarized the following issues that occurred during the test process:

1. **Confusion about professional terminology.** Some terminologies that appear in the prototype can cause the users to fall into confusion. The terminologies that appear in the prototype are k-anonymity, l-diversity, t-closeness, sensitive attributes, quasi-identifiers, utility loss, and privacy loss.

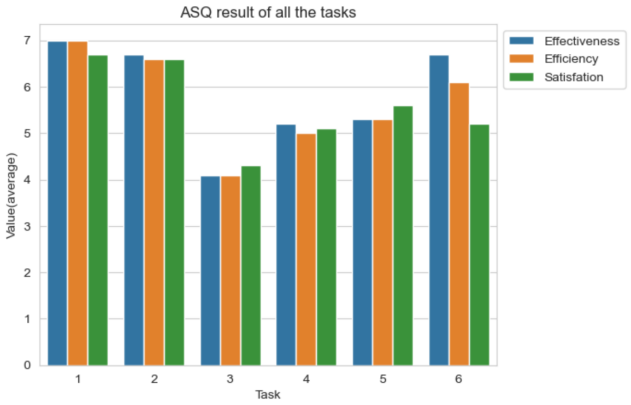


Figure 4.10: ASQ Score of the Tasks

2. **The understanding problem about QIs and SA.** In task 2, we provide an attribute selector to let users select the QIs and SA. We let the user select the QIs and SA separately. Some users thought that they should select all the attributes they wanted to publish first and then select an SA from these attributes. Besides, some users thought that "sensitive attribute" meant the attribute should not be leaked. They are confused about why sensitive attributes will be published.

3. **Response problem of the web page.** Due to the problem in the front-end code, every time the user clicks on the web page, the web page will rerun, which leads to more time consumption and lower user satisfaction.

4. **Interaction problem of the scatter plot.** Users naturally assume that interacting with the scatter plot should directly influence the selection of anonymization parameters, especially the visual representation that implies a connection between data points and parameter choices. This issue might arise due to the low level of the prototype.

5. **Problem in understanding utility loss and privacy loss.** The users can not understand the meaning of these two indicators. We will perhaps add more hints to help the users understand the utility loss and privacy loss.

Figure 4.11 indicates the distribution of the 5 issues that occurred in the first test iteration. The bar chart shows the frequency of these 5 issues. We can see that issue 4 (interaction of the scatter plot) appears most frequently, reaching 80%. Next is issue 1 (confusion about the terminologies) with a frequency of 60%. The occurrence frequencies of issue 3 (response of web pages), issue 2 (QIs and SA selectors), and issue 5 (understanding of utility loss and privacy loss) are 50%, 40%, and 30% respectively. In the second iteration, we will mainly address these issues.

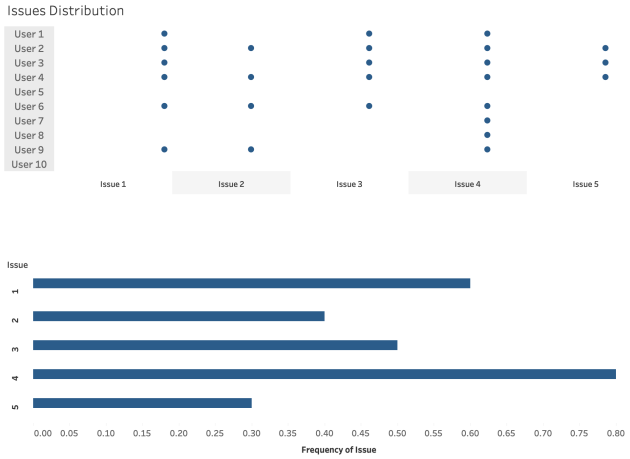


Figure 4.11: Issues distribution in the first test iteration

4.2.2 Improvement

In the second iteration of usability testing, our goal is to assess the impact of the high-fidelity prototype on functionality and user experience, identifying improvements and new challenges. We use a realistic scenario task form the core of this iteration, enabling users to engage with the prototype in practical use cases. Then the After Scenarios Questionnaire (ASQ) provides quantitative feedback on user satisfaction, allowing us to compare results with the low-fidelity model's Task 6, thereby measuring the influence of the high-fidelity model. Besides, qualitative insights are gathered through think-aloud method, encouraging users to vocalize thoughts and feedback during task execution, which helps us get a deeper understanding of users' experiences, unveiling potential challenges and areas for improvement. Furthermore, the validation of identified issues in lo-fi prototype ensures that improvements made in the high-fidelity model are effective while also uncovering any new challenges. Moreover, to offer a comprehensive evaluation, the System Usability Scale (SUS) questionnaire is utilized, providing a holistic assessment of the entire high-fidelity prototype's usability.

Test Participants

Similar to the first usability test, the second iteration involved the recruitment of 10 diverse participants. Notably, these participants differed from those in the first test, comprising individuals from various backgrounds. The participant pool included former high school and college classmates, as well as individuals who have already retired, fostering a broad range of perspectives. This diverse group represented a spectrum of professional experiences, with some participants working as data analysts and others possessing minimal computer usage experience. To ensure a balanced representation, Figure 4.12 illustrates the distribution of participants across age, gender, and experience in data process-

ing. The effort to evenly distribute these basic personal attributes aimed to mitigate potential influences on experimental outcomes.

Figure 4.12: Participants distribution in the second usability test.

Realistic Scenario Task

While the first iteration encompassed six distinct tasks to evaluate various UI components and independent functionalities, the second iteration opted for a consolidated and comprehensive approach. This decision stemmed from the prototype's linear and relatively straightforward workflow, deeming a single task sufficient to encapsulate its entire range of functionalities. This streamlined approach facilitated users in gaining an overarching understanding of the prototype, easing the subsequent evaluation through the System Usability Scale (SUS) questionnaire.

The primary objective of the first iteration was to assess the usability of individual UI components through isolated sub-tests, focusing on each component's usability. In contrast, the second iteration aimed to evaluate the overall usability when integrating all UI components. Although the tasks differed in expression between the two iterations, the fundamental testing goal remained consistent: finding the optimal anonymization strategy using the provided UI components. In addressing RQ2, the first iteration emphasized individual UI components, while the second iteration focused on the aggregated usability of the entire set of UI components.

The selected task was crafted around a realistic scenario involving data anonymization. The primary objective for users was to anonymize a dataset within the context of a genuine use case. The test scenario commenced with a description of the dataset, providing users with the necessary context for the subsequent task. The main focus was on user engagement with key components, including the data uploader, selectors for Quasi-Identifiers (QIs) and Sensitive Attributes (SA), the gauge, spreadsheet, and the scatter plot, as illustrated in Figure 4.13. This strategic design ensured that users interacted with all crucial aspects of the prototype, paving the way for a comprehensive evaluation of its capabilities.

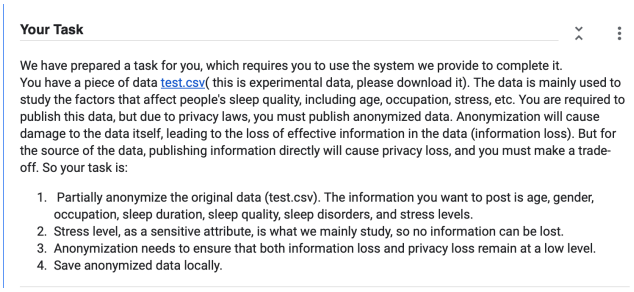


Figure 4.13: Task of the second usability Test.

After Scenario Questionnaire

Following the completion of the designated task, users were prompted to fill out the After Scenario Questionnaire (ASQ). A comparative analysis was conducted by juxtaposing the average ASQ scores of the 10 users in the second usability test with the results obtained from task 6 in the first usability test. These two tasks were considered equivalent as both aimed to identify anonymized data aligning with privacy requirements, with a primary focus on the scatter plot, while other UI components supplemented the task rather than being the central focus (Figure 4.14).

The bar chart illustrates a substantial improvement in user satisfaction and the prototype's efficiency during the second iteration. Enhanced efficiency is attributed to the improved application response speed under the Dash framework and simplified interaction in the scatter plot, streamlining users' efforts in finding optimal anonymization strategies. We speculate that the elevated user satisfaction stems from several factors. Firstly, the high-fidelity prototype alleviated confusion caused by terminologies, eliminating the need for users to grapple with complex terms like "k-anonymity." Secondly, the aesthetically pleasing interface of the high-fidelity prototype contributed to user satisfaction by simplifying the workflow. Lastly, the interaction with the scatter plot positively impacted user satisfaction, compensating for a slight decrease in effectiveness. Despite a marginal dip in effectiveness, the subjective nature of user ratings and the different participant pools between iterations suggest that this difference can be considered negligible.

System Usability Scale

The Average SUS score for the high-fidelity prototype is illustrated in Figure 4.15. To facilitate comparison, the scores were adjusted to a unified scale of 0 to 10, with 10 representing excellent system performance on each corresponding question. The high-fidelity prototype achieved an overall score of 84.75, signifying a commendable usability rating (Figure 2.5).

Observing Figure 4.15, most scores surpass 8, indicating quite positive user perceptions. However, questions 4, 7, and 9 received comparatively lower scores. Question 4 indicates the need for technical support (for instance, termi-

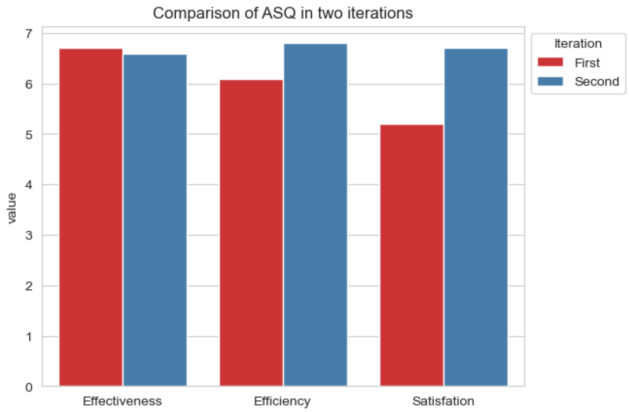


Figure 4.14: ASQ Bar Chart of the First and Second Usability Test

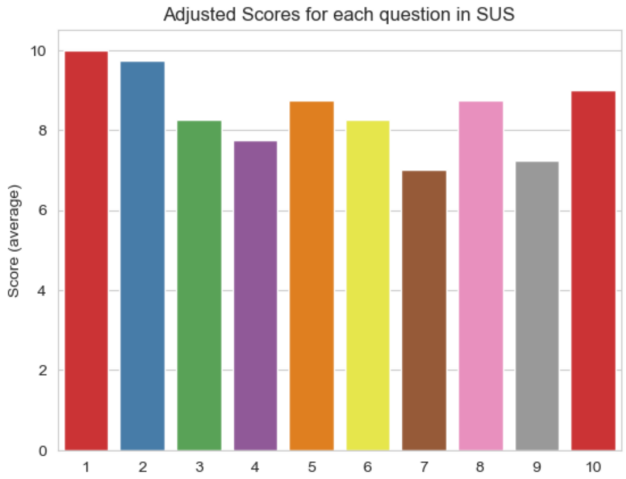


Figure 4.15: SUS Score of the second usability Test. Each column represents a sub-question of the SUS

nology explanation), primarily stemming from users' unclear understanding of sensitive attributes (SA) and confusion about the significance of each point in the scatter plot concerning the selected data attributes. Question 7 evaluates the clarity of feedback and tips, with lower scores attributed to extended wait times during the anonymization process, causing users to miss or overlook some prompts. Lastly, question 9 assesses the perceived difficulty of using various functions, revealing user confusion and dissatisfaction with specific components. (For instance, some users said the interaction with the scatter plot is much more complex than the QIs selector.) A streamlined workflow devoid of ambiguity could further enhance scores in this aspect. We summarized the following issues existing in the high-fidelity prototype:

1. **Understanding of SA**. Although we have removed some terminologies that appeared in the prototype, the SA selector still can't be replaced. However, most users are still confused about the definition of SA. They

asked why it's called a "sensitive attribute". Some data newbies thought selecting a SA was not necessary in the anonymization process.

2. **Understanding of the scatter plot**. When the auto-PCC is done, users can see the scatter plot Figure 3.9(f) appears. However, some users can't get the meaning of the points. They don't understand the correlation between this scatter plot and the previous SA selector. Some of the users didn't try to click the points on the scatter plot even though we already have given the user a tip above the scatter plot.

3. **Feedback from the prototype**. After clicking the button "Run Anonymization" Figure 3.9(d), we provided an information block Figure 3.9(h). Since the anonymization process may take a while, (in some case the process lasts a few minutes.) the users can't get enough feedback from the prototype. They thought the system froze and tried to refresh the website. Furthermore, some users thought the tip above the scatter plot seemed like a title so they didn't expect to get help from this.

The issues identified in the prototype were prevalent among participants, but individuals with varying levels of data experience exhibited different sensitivities to these problems. In other word, their perceptions of the severity of these issues differed. Generally, data newbies expressed a higher level of discomfort with these issues compared to data beginners and experts. For instance, data newbies might struggle with terminologies, hindering them from proceeding, while data beginners only experienced a momentary hesitation upon encountering technical terms, without significantly impeding their ability to take the next steps. Therefore, addressing these issues is deemed crucial for enhancing the usability of the prototype, particularly for users with limited data knowledge. In conclusion, the challenge lies in the inherent complexity of the anonymization process. The term "Sensitive Attribute" remains ambiguous to users, and the necessity of selecting an SA for anonymization is not universally understood.

Some recommendations would be considered in the next iteration. First, implement more intuitive cues, possibly through interactive tutorials or examples, to enhance users' comprehension of SAs and their role in the anonymization process. Second, enhance the tooltip or provide guided interactions to explicitly convey the relationship between the scatter plot and SA selection, encouraging users to interact with the points. Third, implement progress indicators, status updates, or estimated completion times to alleviate user concerns during prolonged processes. Additionally, revise the design or placement of tips to enhance visibility and distinguish them from titles.

According to the feedback from the users, we noticed that some problems that appeared in the lo-fi prototype were solved. In the second usability test, the users didn't mention even once about the meaning of the utility and privacy loss indicator. Moreover, the users didn't complain about the running time of the prototype. They thought for getting the anonymized data, waiting several

minutes was ”acceptable”. The users also mentioned that the response time among user actions is ”not perceptible”. however, the issues of scatter plot's interaction and SA selector still exist in the hi-fi prototype. To address these problems, we will do the third prototyping iteration and a third usability test to validate if these problems have been solved.

4.2.3 Validation

The primary objective of this iteration is to validate the resolution of issues identified in the second iteration of the optimized high-fidelity prototype. The procedure involves a repetition of the realistic scenario task. To quantitatively assess the prototype's performance, a combination of usability metrics, including ASQ and SUS, will be employed. These metrics provide valuable insights into user satisfaction, efficiency, and overall usability, allowing for a comprehensive evaluation of the refined prototype. Furthermore, we focused discussions with users to gather qualitative insights into their expectations and solicit suggestions for future iterations. Direct feedback from users is crucial in guiding the direction of future enhancements. Engaging users in discussions enables us to leverage their expertise, preferences, and expectations, thereby shaping the development roadmap in alignment with user needs.

Test Participants

The participants in the third iteration of testing continue to reflect a diverse range of individuals, maintaining the diversity observed in the previous two iterations. As illustrated in Figure 4.16, the distribution of participants includes variations in age, gender, and experience in data processing. Notably, two participants in their early thirties are professionals in the field of data analysis, and their inclusion is anticipated to bring valuable insights and constructive experiences to the evaluation process. This diverse composition of participants ensures a comprehensive assessment of the prototype's usability, considering perspectives from individuals with varied backgrounds and levels of expertise.

Figure 4.16: Participants distribution in the third usability test

After Scenario Questionnaire

Figure 4.17 depicts the ASQ results for the same task across the three iterations. It is evident that the second iteration witnessed a significant improvement in efficiency and satisfaction compared to the first iteration. However, the ASQ metrics in the third iteration show no significant changes from the second iteration and even exhibit a slight overall decrease in scores. We consider the reason of this result in three aspect: First, introducing new users may lead to subjective differences as their expectations and experiences with the system might differ from previous users, impacting overall user satisfaction evaluations. Second, ASQ, being rooted in user subjective experiences, is susceptible to individual biases. If significant issues were addressed in the second iteration, minor optimizations in the third iteration might not dominate users' subjective evaluations. Third, users typically prioritize aspects that significantly impact their experience and task performance. Minor improvements may have a lower perceptual impact. The sensitivity of ASQ in capturing subtle changes might also be a contributing factor.

In summary, the After Scenario Questionnaire (ASQ) provided a broad overview of the prototype's utility in the third iteration, indicating stable performance without big issues. While ASQ's coarse-grained nature might not capture subtle improvements, the results affirm the prototype's continuous improvement and stability. This outcome emphasized the iterative development process's positive trajectory.

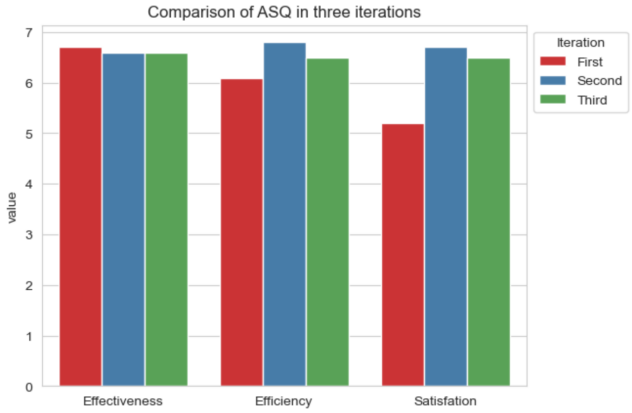


Figure 4.17: Comparison of ASQ in three iterations

System Usability Scale

In the third iteration, the System Usability Scale (SUS) yielded an impressive overall score of 95, reflecting a commendable level of usability (Figure 2.5). As depicted in Figure 4.18, all individual question scores surpassed nine points. Notably, questions 4, 7, and 9, which exhibited lower scores in the second iteration, experienced substantial improvements. The slight enhancements in

scores across other questions further indicate an overall positive trend in user satisfaction and perceived usability. These outcomes affirm the efficacy of refinements made during the iterative design process.

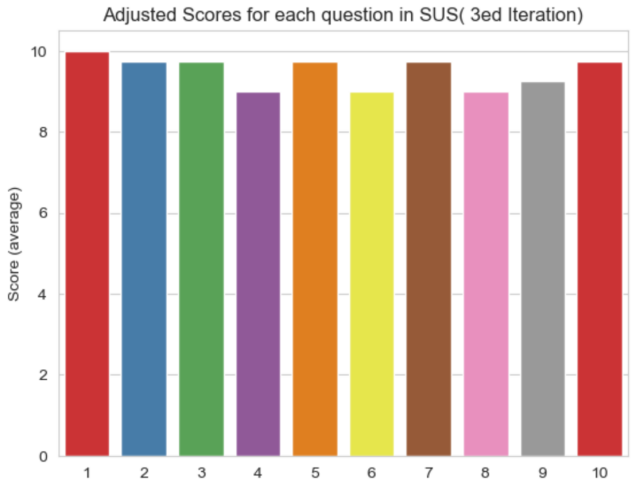


Figure 4.18: SUS Score of the third usability Test

The notable improvement in the third iteration's SUS results can be attributed to several factors. First, the refinement of the prototype based on user feedback from the second iteration played a crucial role. Addressing issues identified in earlier tests likely contributed to a smoother and more user-friendly experience. Additionally, the introduction of new users in the third iteration might have brought a fresh perspective. Moreover, it's essential to acknowledge that the substantial increase in scores, especially for questions 4, 7, and 9, may also reflect the subjective nature of SUS responses. Users did not find issues occurred in the second iteration, influencing their perceptions of the entire prototype's usability.

Suggestions From Participants

While the absolute score comparison between these two iterations might have limitations due to participant variability, it provides a trend indicating the effectiveness of our iterative optimizations. However, the uniformly high scores above 9 raise considerations. Limited participant understanding of our business logic, testing primarily focused on positive experiences, and a single-dimensional testing dataset might not fully capture nuanced functionalities and real-world application scenarios. Besides, The convenience of SUS testing comes with a trade-off between ease of use and evaluation granularity. Although offering intuitive insights, SUS might not delve into the depth of utility and nuanced requirements needed for practical applications. Therefore, while indicative, caution is warranted in extrapolating the absolute success of the prototype from SUS scores alone.

Recognizing the limitations of our usability testing to comprehensively address all challenges in anonymous data publishing across the three iterations, we conducted in-depth discussions with three participants among the 30 testers. These three participants, all professionals in the field of data analysis, exhibited significant interest in our prototype. The discussions delved into future directions for optimizing the prototype, considerations for introducing new functionalities, and exploring potential integrations with other business processes. These conversations aimed to extract valuable insights from experienced users, guiding the roadmap for further improvements and potential applications of the prototype within broader business contexts.

One user emphasized that our current usability test represents a short-term evaluation. They suggested that prolonged usage might unveil new issues and user requirements. The user also questioned whether the prototype, in real-world applications, could genuinely meet the utility and privacy needs as envisioned by users. Recommending a shift towards long-term user testing, the user highlighted the potential for discovering additional functionalities over time. Furthermore, the user pointed out the current prototype's relatively limited scope, citing examples such as its exclusive support for CSV data and only three available anonymization methods. These insights underscore the importance of continuous development and expansion in response to evolving user needs and expectations.

Another user expressed difficulty in comprehending the scatter plot, particularly for users less familiar with data analysis. The user found it challenging to make decisions, especially when scatter plot results were unfavorable. To address this, the user suggested the inclusion of highlighted ideal data points to guide users in making selections. Additionally, the user recommended dynamic usage guides, offering a simple and intuitive understanding of the principles behind anonymization methods. This would empower users to make more informed and rational choices in the anonymization process. These insights emphasize the importance of user-friendly features and educational resources to enhance the overall usability of the prototype.

The final interviewee emphasized the prototype's potential integration into various application domains, highlighting distinct custom requirements in each area. In the realm of business intelligence and data analysis, the prototype offers opportunities for large-scale data analysis without exposing personal customer information, thereby generating valuable business insights. In the education sector, schools and educational institutions can leverage anonymized student data for teaching improvements and academic research. Regarding government data sharing, various levels of government can utilize anonymous data sharing to support urban planning, traffic optimization, and social policy decision-making.

The interviewee envisioned the prototype evolving to provide more customized solutions in the future, catering to specific needs in different domains. Therefore, future development might involve finer customization to ensure op-

timal performance in specific application scenarios and better meet user expectations in particular fields. This perspective underscores the prototype's potential broad applicability across multiple domains and provides valuable guidance for future improvements and developments.

5 Discussion

In addressing RQ1 concerning the development of a tool supporting data controllers in achieving a balanced privacy-utility trade-off during microdata anonymization, our study introduced two innovations. The Auto PCC method, building upon Chou's research[CWM16], automated the exploration of comprehensive anonymization parameter selection strategies. Simultaneously, we proposed a novel approach by integrating privacy and utility considerations quantitatively into a tool, utilizing scatterplots to represent the trade-offs of various anonymization strategies. The case study demonstrated the effectiveness of Auto PCC and scatterplot across diverse datasets, revealing vertical stripe clusters that facilitated the identification of "local optimal points" where privacy and utility were well-balanced. However, challenges arose when dealing with small SA value ranges, indicating a limitation of our method in such scenarios.

The case study further analyzed the effectiveness of various anonymization strategies, dispelling the notion of a one-size-fits-all solution. Unlike traditional heuristic methods, our approach provided users with diverse choices, allowing them to tailor anonymization strategies to their individual needs. This finding underscores the uncertainty inherent in guiding users through anonymization strategy selection using traditional approaches, while our method offers an intuitive way to explore a range of options.

Turning to RQ2, which investigated the usability of interface elements in aiding data controllers to find the optimal utility-privacy trade-off, our usability tests comprised three iterations. In the first iteration, individual UI components (spreadsheet, numerical indicator, scatter plot) were evaluated for their usability in aiding users in making trade-off decisions. While each component exhibited good usability individually, the second iteration, aggregating these components, demonstrated a significant improvement in overall usability. The third iteration involved refining the aggregated UI components, resulting in further enhanced usability.

Contrary to expectations, the single UI components did not exhibit significant differences in usability when guiding users through anonymization strategy selection. However, the aggregation of these components demonstrated a substantial increase in tool usability. This highlights the synergistic effect of combining multiple UI elements in aiding users in making informed decisions about the utility-privacy trade-off.

Despite these contributions, our study has limitations. The datasets chosen for the case study, while diverse, may not encompass all types encountered in real-world scenarios. The simplicity of our usability tests, designed for clarity, sacrificed granularity, and the single testing scenario may not capture all po-

tential issues. Future work should involve long-term, fine-grained evaluations, considering customizations for specific domains.

In summary, our research introduces a novel visualization-supported anonymization strategy selection tool, distinct from previous studies. While not innovating in the quantification of privacy and utility, our integration of visualization provides a unique framework that combines the concern of privacy and utility together. The findings emphasize the importance of considering diverse anonymization strategies and the impact of combining multiple UI elements for improved usability.

6 Conclusion

In this study, our focus was on addressing the trade-off between privacy and utility in data anonymization, realized through the development of the "VisTool" application. By visualizing the privacy and utility of anonymization strategies in one integrated framework, we successfully assisted users in finding a balance that maximizes utility while preserving privacy. Through an in-depth case study, we demonstrated the functionality of VisTool and delved into the parameter selection challenges of the PCC method, offering valuable insights for optimizing anonymization methods.

However, the usability testing process revealed some challenges. Limitations in short-term testing and a singular testing scenario may have led to somewhat one-sided evaluation results. To mitigate such issues, future research should incorporate long-term testing for a more comprehensive understanding of VisTool's real-world performance. Additionally, we identified areas for improvement, particularly in cases where the value domain of the sensitive attribute (SA) is limited. Addressing these issues in future research will enhance VisTool's effectiveness.

The significance of this research lies not only in filling gaps in the current research landscape but also in providing a novel approach to balancing privacy protection and data utility. The successful development and iterative optimization of VisTool lay a robust foundation for future anonymization studies. The program's scalability offers personalized solutions for users in diverse fields such as business intelligence, education, and government data sharing.

We special thanks to Dr. Franzen for the invaluable guidance throughout this thesis. Gratitude is also extended to all peers, teachers, friends, and family who supported and provided valuable insights during the testing process. The success of this study is a collaborative effort, and we look forward to making further breakthroughs in the field of data anonymization in the future.

Bibliography

[AD03] William Albert and Eleri Dixon. Is this what you expected? the use of expectation measures in usability testing. In *Proceedings of the Usability professionals association 2003 Conference, Scottsdale, AZ*, 2003.

[Awa] Abid Ali Awan. Pokémon legendary data. `https://www.kaggle.com/datasets/kingabzpro/pokmon-legendary-data`.

[BA05] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*, pages 217–228. IEEE, 2005.

[Bat] Shubham Bathwal. Flight price prediction. `https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction`.

[BKM08] Aaron Bangor, Philip T Kortum, and James T Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594, 2008.

[BKM09] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

[Bro86] John Brooke. System usability scale (sus): a quick-and-dirty method of system evaluation user information. *Reading, UK: Digital equipment co ltd*, 43:1–7, 1986.

[CWM16] Jia-Kai Chou, Yang Wang, and Kwan-Liu Ma. Privacy preserving event sequence data visualization using a sankey diagram-like representation. In *SIGGRAPH ASIA 2016 symposium on visualization*, pages 1–8, 2016.

[DAS] Dash. `https://dash.plotly.com/`.

[DFT01] Josep Domingo-Ferrer and Vicenc Torra. A quantitative comparison of disclosure control methods for microdata. *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies*, pages 111–134, 2001.

[DL86] George T Duncan and Diane Lambert. Disclosure-limited data dissemination. *Journal of the American statistical association*, 81(393):10–18, 1986.

[DP91] George T Duncan and Robert W Pearson. Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6(3):219–232, 1991.

[DPCTH22] Daniel De Pascale, Giuseppe Cascavilla, Damian A Tamburri, and Willem-Jan Van Den Heuvel. Real-world k-anonymity applications: the\textsc {KGen} approach and its evaluation in fraudulent transactions. *arXiv preprint arXiv:2204.01533*, 2022.

[Duh12] C. Duhigg. How companies learn your secrets, New York Times Magazine, Feb. 16, 2012.

[FBF77] Jerome H Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):209–226, 1977.

[FLA] flask. `https://flask.palletsprojects.com/en/3.0.x/`.

[HIFa] Hi-fi github repository. `https://github.com/yaozheng600/VisTool3.0`.

[HIFb] Hi-fi prototype. `https://vistool3-b3e40b0f9c57.herokuapp.com/`.

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000.

[JWX10] Xuxian Jiang, Xinyuan Wang, and Dongyan Xu. Stealthy malware detection and monitoring through vmm-based "out-of-the-box" semantic view reconstruction. *ACM Transactions on Information and System Security (TISSEC)*, 13(2):1–28, 2010.

[kag] Kaggle. `https://www.kaggle.com/`.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[LDR05] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60, 2005.

[LDR06] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering (ICDE'06)*, pages 25–25. IEEE, 2006.

[Lew91] James R Lewis. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the asq. *ACM Sigchi Bulletin*, 23(1):78–81, 1991.

[LL09] Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526, 2009.

[LLV06] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2006.

[lo-a] lo-fi prototype. `https://vistool.streamlit.app/Exploration`.

[lo-b] lo-fi repository. `https://github.com/yaozheng600/VisTool`.

[LS09] James R Lewis and Jeff Sauro. The factor structure of the system usability scale. In *Human Centered Design: First International Conference, HCD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009 Proceedings 1*, pages 94–103. Springer, 2009.

[LWFP06] Jiuyong Li, Raymond Chi-Wing Wong, Ada Wai-Chee Fu, and Jian Pei. Achieving k-anonymity by clustering in attribute hierarchical structures. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 405–416. Springer, 2006.

[mea] measuringu. `https://measuringu.com/calculators/problem_discovery/`.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[NL93] Jakob Nielsen and Thomas K Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213, 1993.

[Pan] Pandas. `https://pandas.pydata.org/`.

[PLO] Plotly. `https://plotly.com/`.

[Pmo] Pmohun. Complete historical cryptocurrency financial data. https://www.kaggle.com/datasets/philmohun/cryptocurrency-financial-data.

[Sau11] Jeff Sauro. *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC, 2011.

[str] streamlit. https://github.com/streamlit/streamlit.

[TDF03] Vicenç Torra and Josep Domingo-Ferrer. Record linkage methods for multidatabase data mining. *Information fusion in data mining*, pages 101–132, 2003.

[Tel] Telesoho. privacy-preserver. https://github.com/telesoho/privacy-preserver/tree/master/preserver.

[Tha] Laksika Tharmalingam. Sleep health and lifestyle dataset. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset?resource=download.

[WCC+17] Xumeng Wang, Jia-Kai Chou, Wei Chen, Huihua Guan, Wenlong Chen, Tianyi Lao, and Kwan-Liu Ma. A utility-aware visual approach for anonymizing multi-attribute tabular data. *IEEE transactions on visualization and computer graphics*, 24(1):351–360, 2017.

[Wik22] Wikipedia contributors. General data protection regulation — Wikipedia, the free encyclopedia, 2022. [Online; accessed 3-January-2023].

[XT06] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 229–240, 2006.

[XWP+06] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 785–790, 2006.