Perceptions and Practicalities for Private Machine Learning

by

Bailey Kacsmar

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Michelle Mazurek Associate Professor, Department of Computer Science University of Maryland
Supervisor:	Florian Kerschbaum Associate Professor, Cheriton School of Computer Science University of Waterloo
Internal Members:	N. Asokan Professor, Cheriton School of Computer Science University of Waterloo
	Gautam Kamath Assistant Professor, Cheriton School of Computer Science University of Waterloo
Internal-External Member:	Mahesh Tripunitara Professor, Electrical and Computer Engineering University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapter 3 of this thesis was co-authored with Kyle Tilbury, Miti Mazmudar, and Florian Kerschbaum [110]. In particular, the statistical analysis and figures were generated by Kyle Tilbury. Miti Mazmudar and I performed the qualitative coding analysis and all parties collaborated on the study design.

Chapter 4 of this thesis was co-authored with Chelsea H. Komlo, Florian Kerschbaum, and Ian Goldberg [109] Chelsea and I worked collaboratively on the writing and analysis for this work. Ian and I worked on additional writing and the case study analyses included in the appendix. Florian contributed to the threat model development.

Chapter 5 of this thesis was co-authored with Mark R. Thomas, Thomas Humphries, Diwen Zhu, and Florian Kerschbaum. Implementation and experiment execution was done by Mark, Thomas, and Diwen.

Chapter 6 of this thesis was co-authored with Vasisht Duddu, Kyle Tilbury, Blase Ur, and Florian Kerschbaum. Vasisht and I performed all of the data analysis while Kyle and I processed all the interview transcriptions. Florian contributed an early version of the section describing private set intersection and Blase contributed to writing the introduction and background section as well as study design.

All other chapters in this thesis contain original work authored under the supervision of Florian Kerschbaum.

Abstract

Privacy in machine learning holds great promise for enabling organizations to analyze data they and their partners hold while maintaining data subjects' privacy. In this thesis I show that private computation, such as private machine learning, can increase end-users' acceptance of data sharing practices, but not unconditionally. There are many factors that influence end-users' privacy perceptions in this space; including the number of organizations involved and the reciprocity of any data sharing practices. End-users emphasized the importance of detailing the purpose of a computation and clarifying that inputs to private computation are not shared across organizations. End-users also struggled with the notion of protections not being guaranteed 100%, such as in statistical based schemes, thus demonstrating a need for a thorough understanding of the risk form attacks in such applications. When training a machine learning model on private data, it is critical to understand the conditions under which that data can be protected; and when it cannot. For instance, membership inference attacks aim to violate privacy protections by determining whether specific data was used to train a particular machine learning model. Further, the successful transition of private machine learning theoretical research to practical use must account for gaps in achieving these properties that arise due to the realities of concrete implementations, threat models, and use cases; which is not currently the case.

Acknowledgements

I would of course like to thank my advisor, Florian Kerschbaum, for all his input and support throughout my PhD as well as my committee members, N. Asokan, Gautam Kamath, Mahesh Tripunitara, and Michelle Mazurek for their time and insights.

Finally, this thesis would not exist without my amazing colleagues and collaborators in the CrySP lab and outside it. Robert A. Feline (aka Bob, the literal cat) attended so many research meetings and sat beside me while I worked on much of the research in this thesis; especially when the lab was closed for Covid. Kyle has been supportive and calming and the procurer of many espressos throughout this whole process. He also could be counted on to remind me to go outside when I'd been staring at my computer too long and was getting frustrated. Miti and I started together and I cannot overstate how much I appreciate our many walks and talks. Thomas became my co-instructor for our course and we kept each other sane throughout. Thank you to everyone, I'm not good at expressing such things, but I really appreciate all of you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Collaborations and contracts between companies increasingly involve the disclosure of data. For example, MasterCard sold a stockpile of transaction data to Google to track whether Google ran ads that led to a sale at a physical store [21]. Private computation, is a notion of data analysis where there is some shareable output that can only be computed using private unshareable data. Examples include private set intersection [38, 108, 175] and federated machine learning [145, 212] which employ techniques that allow companies to compute over users' data without explicitly disclosing the data to other parties. However, private computation is not a silver bullet to problems associated with sharing user data. User consent is still needed, but cannot be acquired if the users are unable to comprehend the implications of such computations on their data. Although privacy policies should contain information for users about the data a company collects and how that company uses the data, such documents are hard to read and rarely read, making the implications inaccessible to users [141, 161]. To this end, in this thesis, I develop communication mechanisms for users alongside my technical work in private and secure machine learning. Without clear criteria that show what the privacy risks of collaborative learning are, it is not possible for people to make informed decisions about the inclusion of their data in training sets or other datasets. Without knowing what granularity of privacy controls people want, researchers cannot develop the tools that empower them to make informed choices. Thus, this thesis advances systems that enable people to understand what it means for their data to be used in private computation, and specifically private machine learning.

Problem Statement. In this thesis I advance the theory and practice of privacy in machine learning through the following avenues. In addition to attacks on privacy, I

include user perceptions, concerns, and comprehension in designing for privacy in machine learning. I show factors of machine learning and private computation that impact end-users to access and understand the risks of their data being used in machine learning. These factors determine whether people feel safe sharing their data for use in training and what machine learning designs require to give the granularity of control and protection people want over the use of their data. Therefore, in this thesis I investigate the following research question.

> *How can private machine learning be designed with accurate assessments of the privacy risks in a way that can be understood and consented to by the subjects of the data being computed over?*

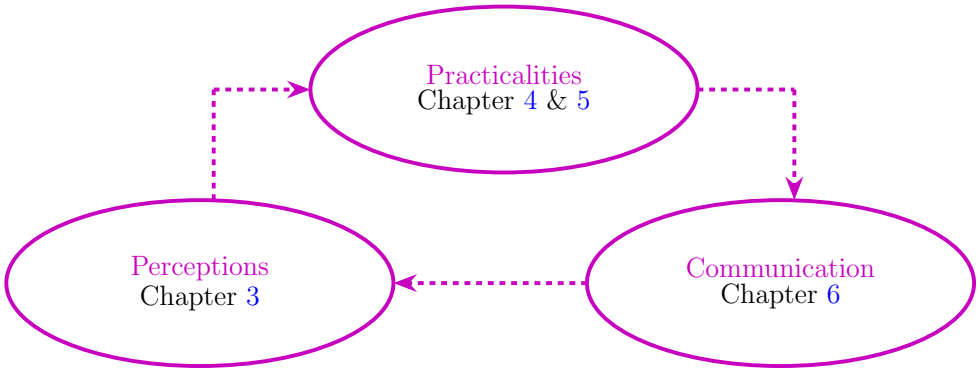


Figure 1.1: Thesis organization

Contributions Although existing technical designs provide some formal guarantees, it is critical to define specific metrics to evaluate privacy guarantees for any specified threat model. That is, metrics are required that show what designs can satisfy participants that have different privacy requirements or trust in the system or organization performing the data analysis. This thesis creates the necessary foundation for such metrics through demonstrating the risks associated with amplifiable existing attacks in different settings (Chapter 5). Further, it shows the level of detail that can lead to variance in end-users privacy preferences (Chapter 3) and how to communicate those factors (Chapter 6). All of these are necessary in order to deploy practical real-world privacy preserving machine learning systems. Including both an understanding of technical risks and human-factors, this thesis develops improved technical systems that address otherwise neglected areas; in the form of a proactive verifiable secret sharing scheme (Chapter 4).

Chapter 2

Background

Privacy is not limited to one definition or concept. In this section I highlight technical notions of privacy, conceptual privacy, and legal privacy. Further, there is the notion of usability and accessibility as it specifically applies to privacy tools and private computation. Finally, we can consider a specific form of private computation, privacy in machine learning, and see how it encompasses significant breadth in terms of privacy.

2.1 Technical Privacy

Technical definitions of privacy, generally speaking, include two key aspects; what is being protected, and from who. Within technical privacy, privacy guarantees are protections that are achieved given a series of assumptions are met. These assumptions may be about the potential adversaries, system complexities, or statistics.

For example, consider the following two types of adversaries. An *honest-but-curious* (HBC) adversary will not deviate from the agreed upon behaviour within a protocol or system, but will try to learn as much information as they can within the defined bounds. An HBC adversary will use any information exposed to them, as well as potentially collude with others in an effort to learn additional private information. A *malicious* adversary is not bound to any expectation of behaviour (though cannot break for instance ability assumptions such as computational limitations), and they can participate both honestly and dishonestly at will. A malicious adversary can impersonate others, elect to not participate, or participate disruptively.

An adversary attacking a privacy tool can have different goals and assumed behaviours. While preventing the availability or use can be the goal of an attacker, these goals will generally not impact privacy.

Learning private information. The information targeted by an adversary will vary depending on the setting. In threshold schemes (e.g., Chapter 4), the secret information is the shares, which are used to recover the secret. In private machine learning (e.g., Chapter 2.5), the secret information may be the training data, which an adversary can attempt to learn through membership inference attacks [104].

Modifying information. An adversary may wish to modify information without detection. Doing such a modification may allow an adversary to better perform an attack to learn secret information. However, modifying the secret information, perhaps via poisoning a machine learning model, will not directly reveal private information such as training data [79].

Differential Privacy. In the last fifteen years, differential privacy is one definition that has risen to prominence in terms of use for achieving technical privacy [60]. Conceptually, differential privacy is used to achieve a technical privacy guarantee. Essentially, the guarantee is that plaintext data can be contributed to a dataset without any individuals' data significantly changing the outputs of select queries or computations over the dataset. More formally, there are different definitions of differential privacy that are applied depending on what computations over the dataset are desired. One definition is ϵ-differential privacy.

Definition 1 *(ϵ-Differential Privacy [60]). A randomized mechanism $M : \mathcal{D} \mapsto \mathcal{F}$ provides ϵ-differential privacy (ϵ-DP) iff for all neighbouring inputs D, $D' \in \mathcal{D}$, ie., differing in one element, and all subsets $F \subseteq \mathcal{F}$,*

$$Pr[M(D) \in F] \leq e^{\epsilon} Pr[M(D') \in F],$$

where the probability space is M's coin tosses.

That is, when using ϵ-Differential Privacy [60], the probability of the perturbed output of the mechanism M occurring is essentially the same whether the input was D or D'.

Homomorphic Encryption. While differential privacy provides a technical notion of privacy for computations when plaintext data has been contributed, a different privacy guarantee can be achieved through the use of homomorphic encryption [245]. Homomorphic encryption does not require plaintext values to be shared to achieve the computed

results. In homomorphic encryption, plaintext data is encrypted such that when select computations are performed over the encrypted texts, the decryption of the computed result is the same as the computed result would be over the plaintext values.

Private Computation. Private computation defines an input, an output, and a function with limitations as to what can and cannot be inferred by an adversary, even if the adversary possesses some subset of the input, output. The function enforces the inference limitations through the use of mathematics such as cryptography or statistical guarantees. Private computation may employ differential privacy, homomorphic encryption, as well as other cryptographic techniques to achieve these guarantees. Further, such computations may be between two or more parties, and may involve trusted third parties. Consider the following example of a type of private computation.

Private Set Intersection (PSI). Two or more parties can compute the intersection of their data without revealing data they possess outside of the intersection using private set intersection (PSI). If the contents of the intersection also has some privacy requirements due to the sensitive nature of the data, differential privacy can be used for additional privacy guarantees [108].

2.2 Theories of Privacy

Theories of privacy attempt to define and describe privacy or privacy behaviours [7, 124, 158, 173, 231]. Some theories, such as the Westin categories of 'the fundamentalist', 'the pragmatist', and 'the unconcerned' are used to classify general privacy attitudes, but do not necessarily reflect privacy actions when individuals are presented with specific scenarios [234]. Others, such as contextual integrity focus on an individuals' right to privacy with respect to information about themselves, and to exercise that right differently in different contexts [158]. Information about individuals may be collected by employers, government entities, and friends. Which of these collectors original receives the information is one component of the 'context' or social domains in which information is shared. Once the information is in a different context, whether via use or disclosure, it can no longer be assumed to meet privacy expectations. While theories of privacy do not prescribe how to develop effective tools and technologies, they aid in the interpretation of user actions, behaviours, and needs.

2.3 Legal Privacy

Legal notions of privacy are primarily framed in terms of individual protections from government and from corporations; with legal and financial penalties for non-compliance. Canada has PIPEDA, the Personal Information Protection and Electronic Documents Act, which protects individuals from the collection, use, and disclosure of their data by corporations without the individuals consent [164]. After being updated in 2018, non-compliance with PIPEDA regulations can lead to corporations facing fines up to $100,000. The United States has the Children's Online Privacy Protection Rule (COPPA)[1], the Health Insurance Portability and Accountability Act (HIPAA) [3], and recently the state specific California Consumer Privacy Act (CCPA[2]). Members of the European Union have the General Data Protection Regulation (GDPR) [230], which has stricter requirements and more costly penalties for non-compliance (up to $20 million Euros or 4% of total global revenue). Such legal regulations may impact individuals' perceptions of privacy, and can determine current compliance requirements for privacy. However, changing laws takes time while new technologies are in constant development, and thus these laws cannot encapsulate current and future privacy requirements and expectations.

2.4 Usable Privacy

Privacy tools may require additional user effort over non-privacy preserving tools and therefore require clear motivation before users will choose to use them. This is particularly true when the non-privacy preserving tools only inform users of privacy implications through privacy policies; which are rarely read nor understood by users [141, 161].

Unlike direct to consumer applications, (e.g. Signal[3]), private computation often uses data about users without directly engaging with them. The data subjects, or data donors, provide data to one entity, and then that entity uses the data towards some goal. In a recent study from Agrawal et al. [6], technical expert participants acknowledged the significance of end users. However, they did not consider end user understanding and consent a priority. Despite the indirectness, consent and communication, in terms of how user data is used, is still needed and potentially legally required, although ensuring it is somewhat more complex. Past efforts at communicating implications of private computation or the use of

[1] https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule

[2] https://oag.ca.gov/privacy/ccpa

[3] https://signal.org/en/

technical privacy find that end users struggle to understand technical privacy approaches and desire better explanations [51, 237]. In addition to understanding, past descriptions have struggled to inspire trust and confidence [26, 178]

To effectively design privacy tools that users will feel encouraged to use, it is necessary to study users awareness, understanding, and motivations [11, 51, 160, 188, 220]. While usability can include efficiency and practicality from a technical standpoint, private computation must inspire trust and match the expectations of the data subjects to ensure their continued consent to the use of their data in such computations.

2.5 Privacy in Machine Learning

Privacy-preserving machine learning has the potential to balance individuals' privacy rights and companies' economic goals. However, such technologies must inspire trust and communicate how they can match the expectations of the subjects of the data. In this section, I discuss the breadth of privacy vectors for machine learning and their relationship to user perspectives of the space.

2.5.1 Data Sharing in Machine Learning

In a computation, such as machine learning, there are two main roles that can be held by an individual or a corporation.

Definition 2 *A data subject is an entity whose data is present in the data set being computed over (e.g., the training set) and the data describes the subject or their attributes.*

Definition 3 *A data owner is an entity that holds a dataset that is being contributed to the data analysis (e.g, towards training a shared model).*

In some cases, the data subjects may be the same as the data owner. For instance, this occurs in the case where keyboard users agree to share their typing data to improve keyboard suggestions [239]. In other cases, however, the data subject could be a medical patient and the data owner a medical research lab with a dataset containing information on a number of data subjects. Data owners who are not data subjects may have different privacy expectations or requirements than in the case where the data subjects are the data

owners contributing their data. The situation can be more complicated when the data subject did not provide their data directly to the current data owners. In a setting where the data subject initially gives their data to an organization that becomes the new data owner, the data subject may not have understood their data could be used in collaborative learning, or any corresponding privacy risks [230].

Further adding to the complexity, machine learning exists in many forms, resulting in varying structures that we must account for if we are to communicate the implications to the data subjects. Across "forms", each have their own sub-forms that have unique configurations for training and tasks. In this thesis, we will not be focusing on the types of tasks that can be performed (classification, inference, clustering, etc.). Instead, we will be focusing on the ways sensitive data can be defined in machine learning, how it can be trained over, and what unique "sharing structures" can occur in machine learning.

When we are concerned with data sharing structures, we are referring to cases where either (i) the data subjects are not the data owners training a model, (ii) there are many data owners each contributing data from many data subjects, (iii) there are many data subjects each providing their own data, or there could be some combination of these three cases. Each of these four cases can occur differently depending on the form of machine learning being performed.

Within the field of machine learning, consider the following high-level forms of machine learning. These forms are based on how the "training" data is accessed; with the most conventional of these forms being *stand alone machine learning*.

Definition 4 *Stand-alone machine learning, or centralized machine learning, is any machine learning algorithm (neural network, decision tree, etc.) where all of the data being trained on is located in one place.*

Within stand-alone machine learning, it is possible for each of the three cases (i-iii) to occur. Case (i) is rather typical, and includes things like an email provider training over multiple users' email inboxes. Such an email provider can centralize all the data they want to train over as the sole data owner; but the data subjects are many, each able to have their own expectations for privacy and how their data can be used by the organization providing them a service. Case (ii) can occur as a stand-alone training instance when all of the data owners trust one another. While they may possess data they have collected from each of their own client sets, they are in a scenario where they can deposit all of their collective data in one location for training over. Similarly, Case (iii) can occur in instances where a number of data subjects, each holding their own data, elect to send their data

to a trusted third party controlled centralized location. This can only be done if each of these individuals trust the third-party entity, otherwise, for both Case (ii) and (iii), it is necessary to move to the next form of machine learning; collaborative machine learning.

Definition 5 *Collaborative Machine learning is any machine learning algorithm where the training data is not all held in a centralized location. This includes horizontal federated learning [1], vertical federated learning [105, 236], and any other similarly distributed machine learning.*

Collaborative learning is typically performed when the data to be trained over cannot be placed in a centralized repository. This could be because the collective amount of data is too large too store, there are too many contributors to coordinate all of the data be available at once, or because the contributors do not have a trusted entity they are all willing to provide their data to. Examples for our three cases are as follows. First, for case (i) two health research companies may want to train a model over their respective datasets; but not wanting to send that data directly to one another or to a third party. Note that this solution is not necessarily going to achieve their privacy goals if they also need to guarantee that neither one is able to learn about their training sets (see attacks section below). In the collaborative setting; there is no salient difference between case (i) and case (ii), however, for case (iii) there is another example. Case (iii) includes a technology company that wants to train their typing prediction software for their custom keyboard. In this case, if the users of the keyboard are able to self-enrol contributing their data to the model, they are then fulfilling both the role of data subject and data owner. When performed in a collaborative setting, there is no trusted party who receives the data directly; rather the protocol will execute such that each party provides contributions to the model in the form of a "model update". The nature of such updates depends on the specific implementation used; such as model parameter updates, gradient updates, or some other attribute.

2.5.2 Privacy Protection in Machine Learning

In private machine learning, what needs to be protected can include the training data, the model (e.g., the parameters), and the model outputs or inferences. Attacks on machine learning models aim to learn these sensitive attributes, thwarting their protections. Such attacks include inference attacks, model inversion attacks, and others [22, 91, 94, 168]. Each of these attacks pose risks depending on the context of the data being used and what needs to be protected. Generally, the goal of a membership inference attack is to

determine whether a target data element belongs to the training set that was used to train the machine learning model [207]. If, for example, whether an element was in a dataset is itself sensitive, membership inference attacks can be the most damaging attacks on the privacy of machine learning.

Thus, most technical protection mechanisms focus on one of these aspects, such as protecting the training data, and typically do not secure all of these potentially sensitive attributes. There are a number of protocol designs that perform differentially private machine learning in the non-distributed (or stand-alone) setting [1, 37, 104]. There are also a few designs for applying differential privacy in the collaborative, or federated, setting [85, 205]. Designs for privacy-preserving training can apply combined techniques, including some selection of differential privacy [85], third parties [167], and cryptographic computation [23]. Some constructions employ multiparty computation or homomorphic encryption to achieve stronger privacy guarantees at a high cost to efficiency [225]. One such construction improves upon efficiency using a combined differential privacy and multiparty computation approach [238].

Chapter 3

Perceptions

This chapter is adapted from work that previously appeared as "Caring About Sharing: User Perceptions of Multi-party Data Sharing" at the 2022 USENIX Security Symposium [110]. This chapter shows how there are many factors influencing privacy perceptions; even when considered quite broadly. Within this chapter the data sharing practices covered encompass some of the breadth of structures we highlighted in the previous chapter with respect to the forms of machine learning. That is, given this work we show that the many forms of machine learning, each with their own structures, cannot assume to fit existing understandings of privact perceptions and preferences.

3.1 Introduction

Collaborations and contracts between companies increasingly involve the disclosure of data. Mastercard sold a stockpile of transaction data to Google to track whether Google ran ads that led to a sale at a physical store [21]. Data moving between companies is not limited to direct sales or targeted advertising. Data sharing can also occur through the purchase or merging of companies such as Google purchasing Fitbit [82]. Although Google's purchase of Fitbit includes a statement that the health and wellness data will not be used for Google advertising, it does not clarify how other data could be used and whether the health and wellness data can be used in ways not related to advertising. Through a legal request one user determined Tim Hortons' loyalty program app shared its users' precise location regularly with a third-party (Radar Labs Inc.) that identified users' home, work, travel destinations, as well as visits to a competitor. The third-party ultimately

shared the users' precise locations with Tim Hortons' parent company, Restaurants Brand International [144].

There are even collaborations between technology and health companies that can and do occur. There are collaborations between Google and Ascension [195], Microsoft and Providence StJoseph Health [35], and COVID-19 contact tracing tools [90]. Some of these collaborations only include the use of services, but others require sharing data in some form to perform computations, including machine learning. In addition to these forms of collaborations, the line dividing health and technology companies is blurring with the development of new services such as Amazon Care[1] and Telus Health[2]. Amazon's health care service specifies that patient information is exclusively used for supporting Care Medical, however, it is unclear how this could affect users' understanding and perceptions of health care data being used by technology companies.

We refer to companies that acquire or share data in these ways as collaborating for multiparty data sharing. Mechanisms to perform privacy-enhanced multiparty data sharing exist in the literature as secure computation, such as private set intersection [38, 175] and federated machine learning [145, 212]. While companies, such as Microsoft and Google, may choose to use privacy-enhanced computation in their collaborations, how to convey these practices fairly to users and indeed how users feel about enhanced computations is a question we address within this paper. Multiparty data sharing can be one-way, where only one of the companies in the exchange acquires data, two-way where the parties involved pool their collective data, or an exchange involving more than two-parties.

Although privacy policies should contain information for users about the data a company collects and how that company uses the data, such documents are hard to read and rarely read, making them inaccessible to users [141, 161]. Users who trust one company with their data may not understand that their data could be shared or purchased nor the corresponding privacy risks. However, it can be confusing for people reading privacy policies about sharing their data to understand what their data will be used for and make informed decisions based on their perceptions of it.

Research Questions. We study users' perceptions of multiparty data sharing via an online survey. We analyze users' perceptions of various data sharing events (termed as scenarios), what potential controls users want, and identify avenues for improving regulations and engineering better systems to meet those needs. To this end we address the following research questions (*with salient results emphasized*):

[1]Available across the United States, https://amazon.care/

[2]Manages Canadian health care records, https://www.telus.com/en/health

RQ1: How does the overall acceptability vary across different types of multiparty collaborations? How do the types of companies involved further impact it?

> *The overall acceptability of multiparty data sharing is lower for collaborations that are not reciprocal. The inclusion of a health company in non-reciprocal collaborations is even less acceptable. (Section 3.4.2).*

RQ2: How does acceptability vary in multiparty data sharing for different user controls (consent, purpose, retention)?

> *Across user controls, preferences for consent vary the most between collaboration types, however, opt-in consent is, generally speaking, the most acceptable. (Section 3.4.1 and 3.4.2)*

3.2 Related Work

Privacy Perceptions. Users' perceptions of privacy have shown many changes over the years and so have their preferences [9, 48, 120]. Past work has often focused on data sharing for advertising purposes [46, 140, 228, 242], with the additions of privacy perceptions for IoT, mobile, and smart homes in more recent years [10, 69, 128, 153, 221, 233]. Regardless of whether the data is shared intentionally or unintentionally leaked via a data breach, user perceptions tend to perceive such treatments of their data negatively [71, 111, 139, 140, 180, 204].

Even when a users' data is only disclosed to a single company, different contexts influence what trade-offs users are willing to make at the expense of their privacy in terms of benefits, or how their data is being used [14, 18, 20, 57, 158, 227]. Further complicating matters are 'third parties' or 'partners' that data can be shared with. Users do not understand what these third parties are and how their data can be shared with these parties [180]. In cases where such terms are used in a privacy policy, it can remain ambiguous to users as to who their data can be shared with and thus prevent them from making an informed decision [66, 130].

In general, survey methodology research cautions that respondents may have difficulty predicting their behaviour or be inclined to report the perceived desirable response [177]. In the case of security research, recent work from Redmiles et al. [185, 186] shows that

surveys can provide meaningful results for general constructs. We use a similar survey design to previous work on acceptability for IoT and data breaches [10, 113].

Thus far, research has primarily treated third-parties or partners in much the same manner as privacy policies do. Third parties are treated as monolithic black-box entities that can take many forms and treat data in different ways. Ebert et al. [62] include 'data sharing' among the legal principles of their study, but again it is left as a general concept. In this work, we build on past investigations into user perceptions of data sharing by specifically providing respondents with scenarios based on real-world examples of how their data could be shared with one or more other parties. We revisit whether policy and design decisions relating to these continually evolving multiparty data sharing scenarios can rely on past results, or whether different structures of data sharing result in different perceptions that need to be addressed.

User Controls and Accessibility. Though not strictly targeting the multiparty data sharing setting, methods to provide users with controls include toggles [89], permission settings [117, 129], and privacy nudges [5]. Despite this, such controls can still be hard for users to understand and use [4, 5, 17, 65, 69, 88, 187]. Difficulties associated with providing users with controls to set their own privacy preferences are not limited to the design of such controls. That is, users can be manipulated or tricked such that opting out of behavioural based advertising is limited [88, 127]. With this in mind, we specify explicitly details users may want to have user controls for in the survey. These aspects for potential controls include what purposes users find acceptable for their data, how they want to be informed (to get consent), and how long they will permit their data to be used in this way.

Park and Sandhu proposed *usage control* to generalize these controls and the idea that beyond privacy policies for all users there can be individual controls required for each user [169, 170]. Ebert et al. [62] referred to usage control variables such as storage and retention as legal preferences in their analysis. They do not focus on types of data sharing, but instead on the effect of the contexts of a fitness tracker versus a rewards card. Similar to Park and Sandhu's application to social media controls, in the case of multiparty data sharing, there are many potential parties that users may or may not want to share their data with and the type of data they are willing to share may vary for different companies [169].

Law and Policy. There are a number of regulations, both old and more recent, that apply to the privacy of users' data [47, 159, 164, 190]. However, they do not necessarily provide protections for all of the possible treatments of users' data [135, 162]. Even with

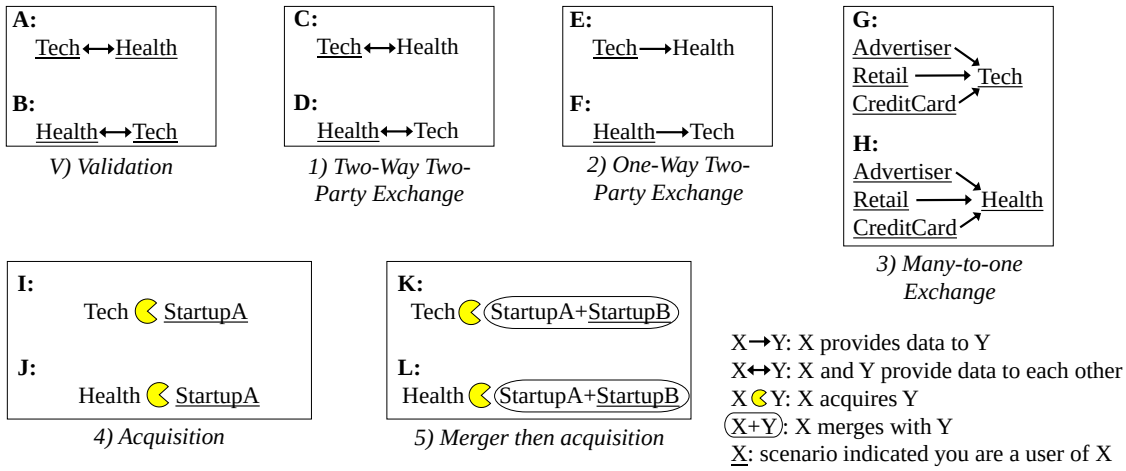


Figure 3.1: Overview of scenarios (A-L) presented in our survey and collaboration types (V, 1-5) that we investigate. For reference, Scenario C, "TechForYou is a large internet company that offers a search engine, email accounts and smartphone platforms to users. GoodHealth runs a chain of hospitals across the country and stores health data for millions of patients during its day-to-day operations. TechForYou and GoodHealth will share the customer data they hold with one another. You are a customer of TechForYou".

the recent California Consumer Privacy Act (CCPA[3]), the right to opt-out of data sales does not stop companies from manipulating users such that it is difficult or unappealing to opt-out[162]. Furthermore, it can only prevent companies from *selling* users' data, it does not prevent companies from sharing or exchanging data with other companies or affiliates. Multiparty data sharing needs to be better understood with respect to user preferences and perceptions to produce more specific regulations addressing all types of collaborations.

3.3 Methodology

We collected 1025 responses to our online survey through SurveyMonkey in March 2021. Each participant was compensated $3.04 for their response and spent, on average, four minutes to complete the survey. Our final participant set is $N = 916$ after excluding the 109 respondents that failed an attention checking question. Respondents could exit the survey at any time and could skip any question in the survey. Our study received

[3] https://oag.ca.gov/privacy/ccpa

ethics approval from our institution's office of research ethics (ORE). See survey at https://bkacsmar.github.io/files/SurveyUsenix2022.pdf.

3.3.1 Survey Design

Prior to the final survey, we ran a pilot study with $N = 26$ participants. We asked participants in the pilot study what they *would* agree to in a multiparty data sharing setting. The pilot had one scenario, between a technology company and a financial institution, to introduce the concept of multiparty data sharing. We used a free-form text response question to gather participants initial thoughts on this scenario and what could influence their perceptions. Our pilot study free-form responses report a desire for user controls that we incorporate into our final survey.

3.3.2 Survey Structure

Each survey provides one of twelve scenarios to respondents followed by a series of questions on user controls and privacy mechanisms. The twelve scenarios are categorized by the number of companies and which companies send and receive data (see Figure 3.1 for an overview of the scenarios). Each collaboration scenario is based upon real-world examples from Canada and the United States. For each question, excluding the free-form responses and correctness checks, respondents select a value from a five-point semantic differential [165] acceptability scale: "Completely Unacceptable", "Somewhat Unacceptable", "Neutral", "Somewhat Acceptable", and "Completely Acceptable" as in Apthorpe et al. [10]. Respondents rate acceptability given specified variables (shown as (a) through (k) in Table 3.5). For analysis, the values we assign to our scale are 1-5 where 1 is "Completely Unacceptable" and 5 is "Completely Acceptable".

3.3.3 Nature of Collaboration

The nature, or type, of the collaboration encodes the number of participating companies and how the data flows between those companies. Notably, we test the inclusion of a health company versus a technology company within the collaboration types. To check whether the ordering of the companies influences respondents, we include two identical scenarios, Scenarios A and B, where the only difference between them is the order in which the health and technology company are introduced. The following defines our five collaboration types with examples.

Between collaboration types. We perform a Kruskal-Wallis test on the distribution of acceptability of each collaboration type (1-5) for each variable ((a) through (k)) and report those with significant differences in Table 3.1. We perform a post-hoc analysis for variables that have significant differences from the Kruskal-Wallis test to identify which collaboration types have pairwise differences. We use Dunn's multiple comparison procedure and show the results in Table 3.2. Only the collaboration type pairs that have significantly different distributions of acceptability are reported. The difference in mean rank (e.g., the mean rank of Type X subtract the mean rank of Type Y) shows the direction of the difference in acceptability collaboration types.

Between collaboration types, the acceptability distribution of...	**Test Statistic**	***p***
...(a) is the same	26.724	<0.001
...(c) is the same	15.113	0.004
...(d) is the same	10.340	0.035
...(e) is the same	12.058	0.017
...(h) is the same	13.261	0.010
...(k) is the same	10.337	0.035

Table 3.1: Kruskal-Wallis test results for the distribution of acceptability of variables between sharing types {1 ($N = 140$), 2 ($N = 150$), 3 ($N = 134$), 4 ($N = 162$), 5 ($N = 170$)} for which the acceptability of the variable differs significantly between data sharing types.

Within sharing types. Each sharing type (1-5) is comprised of two scenarios, so within each type we perform a Mann-Whitney U test for each variable ((a) through (k)). For 'two-way two-party exchange' (type 1), we fail to identify any significant differences in the distribution of acceptability for its constituent scenarios C ($N = 73$) and D ($N = 67$). In 'one-way two-party exchange' (type 2), we identify significant differences between scenarios E and F in seven variables which can be seen in Table 3.3. For 'many-to-one exchange' (type 3), we identify one significant difference between scenario G ($N = 64$) and H ($N = 70$) for 'assumed consent' (variable (c), $p = 0.035$, std. test statistic= -2.107, mean rank difference= 13.84). For 'acquisition' (type 4), we identify a significant difference for 'opt-in consent' (variable (e)) between scenarios I ($N = 79$) and J ($N = 83$) ($p = 0.004$, std. test statistic= -2.915, mean rank difference= 20.24). For 'merger then acquisition' (type 5), we fail to identity any significant differences in acceptability of variables for scenario K

Collaboration Type X, Type Y	Difference in Mean Rank	Std. Test Statistic	p
(a) All scenarios (general)			
2, 4	-75.46	-3.124	0.018
2, 5	-69.42	-2.907	0.037
3, 4	-104.31	4.190	<0.001
3, 5	-98.27	3.990	0.001
(c) Assumed consent			
2, 4	-68.28	-2.825	0.047
2, 5	-68.23	-2.855	0.043
(d) Opt-out consent No pairwise differences due to Bonferroni correction.			
(e) Opt-in consent No pairwise differences due to Bonferroni correction.			
(h) Retained for set time			
2, 4	-71.96	-2.973	0.030
(k) Improving services			
2, 5	-70.38	-2.948	0.032

Table 3.2: Dunn's multiple comparison test results for the distribution of acceptability compared pairwise between collaboration types. All p values are adjusted for multiple comparisons (10 comparisons per variable).

($N = 74$) compared with L ($N = 96$).

Two-way, Two-party Exchange (Type 1). In a 'two-way two-party exchange' there are two participating companies. During the exchange, the two companies send data to and receive data from one another. Four of our scenarios are a 'two-way two-party exchange' (Scenarios A-D). We use two of these four scenarios (C and D) in our collaboration type analysis, and we use the remaining two (A and B) for validation only. Examples of such a collaboration would be two companies that perform a computation, such as private set intersection dual execution, that uses extended methods to ensure both companies receive the result [148].

Within One-Way Two-Party Exchange (E, F), the acceptability distribution of...	**Difference in Mean Rank**	**Std. Test Statistic**	***p***
...(a) is the same	16.04	-2.322	0.020
...(e) is the same	17.47	-2.550	0.011
...(g) is the same	16.11	-2.315	0.021
...(h) is the same	15.19	-2.188	0.029
...(i) is the same	17.22	-2.603	0.009
...(j) is the same	22.22	-3.202	0.001
...(k) is the same	15.24	-2.196	0.028

Table 3.3: Mann-Whitney U test results for the One-Way Two-Party Exchange (collaboration type 2) scenarios {E ($N = 81$), F ($N = 69$)}.

One-way, Two-party Exchange (Type 2). Perhaps the most conventional and well understood collaboration type is the 'one-way two-party exchange' (Scenarios E and F). In this case there are two companies where one acquires data from the other, perhaps in exchange for a monetary amount. Such collaborations could be two parties computing the intersection of data they hold where one party receive the resulting intersection [21]. Other examples of this collaboration type include insurance telematics (use-based insurance) [126] and computing joint cyber threats [28].

Many-to-One Exchange (Type 3). A company may acquire data related to their users from multiple other companies or data brokers. We include two scenarios of this form (Scenarios G and H) with a total of four participating companies. In these 'many-to-one' scenarios, three of the companies are providing data to one other company. This structure in practice, could of course take many forms depending on the number of participating companies and which companies provide or receive data. We chose this structure based on the real-world examples of companies acquiring data from a series of other 'partner' companies. For example, advertising networks may acquire data from any number of sources, including other apps, websites, and their competitors, depending on users' permission settings [67, 119].

Acquisition (Type 4). In our 'acquisition' scenarios, a single party purchases, or acquires, another (Scenarios I and J). Examples of acquisitions relating to data sharing include Google acquiring Fitbit [82], Microsoft acquiring LinkedIn [36], and WealthSimple acquiring SimpleTax [93]. The company SimpleTax promised to never sell its users' data,

however, this did not account for when the company itself was sold. In such acquisitions the data held by a company may be included in its assets and upon purchase becomes available to the acquiring company depending on the applicability of regulations such as the FTC Act [4]. In the case of the purchase of SimpleTax, the explicit promise to never share its users' data was removed from its privacy policy going forward (only affecting data since the purchase) [93].

Merger then Acquisition (Type 5). Generally speaking, the difference between a merger and an acquisition can be thought of as two companies equally choosing to come together as one company in a merger versus one company taking ownership of another during an acquisition. In both cases, assets, which may include data, are consolidated in some manner. We include a scenario where two startups merge, forming a new company, which is then acquired by a third company (Scenarios K and L). In this case it is possible for an individual to have shared their data with one of the original start-ups, with no expectation that these two additional companies they have no connection with would come to possess it. Sometimes a merger with other acquired companies can be a part of an acquisition, and sometimes they are separate events; but they are both possible outcomes for smaller companies [224].

3.3.4 User Controls

Usage control enforcement mechanisms are components that can be written into designs or regulations which give users the ability to specifically set what they agree to. We use eleven usage control variables (listed in Table 3.5 as (b) through (k)) within our survey. The variables are selected from responses to our pilot study and real-world examples. We investigate how purpose of use, data retention, and the method of acquiring consent or notifying users can impact the acceptability of multiparty data sharing scenarios.

Purpose. There are three purposes of data sharing in our survey. These purposes are 'generating advertising revenue', 'providing users with a monetary reward' (e.g., free service, reduced rate [126], or gift-card), and 'improving services' [65, 201]. Note that while we included a variety of examples within the monetary return question, these examples may not have been viewed the same by all respondents. That is, respondents may have interpreted free service as an advertising funded service rather than an additional bonus

[4] https://www.ftc.gov/about-ftc/what-we-do/enforcement-authority

service. Respondents that interpreted a free service in such a way may have been less inclined to consider the service as a monetary benefit in the same sense as a gift card or discount.

Data Retention. Users are known to have misconceptions about what happens when their data is deleted[152]. To prevent misconceptions, our data retention questions provide an explicit duration for each of the three retention questions. The duration values include keeping the data 'indefinitely', keeping the data for a 'specified duration' of time (e.g., three months, one year, etc.), or more ambiguously, keeping the data until the company (or companies) is 'finished using it'. We note here that the deliberate inclusion of the more ambiguous 'after they finished using it' does leave the potential for respondents to interpret it differently. Data may be used by companies in computations such as aggregate statistics, private set intersection, or to train machine learning models. Respondents may differ in whether they believe that continuing to use computations on data means that a company continues to use the data. We left interpretation of when the use ends open to the respondents.

Notification and Consent. We avoid directly asking participants whether they would consent, which would likely be influenced by perceived socially desirable behaviour [121]. Instead, we focus on notification strategies that inform users. Depending on local laws and regulations companies use a variety of methods to inform (or not inform) users how personal data can be used. We select a subset of those methods to evaluate any potential influence on the acceptability of multiparty data sharing.

In our survey we include four questions relating to informing users. First, 'concealed consent', where no formal notification is provided, and the respondents learns of the collaboration via the media. Second, there is 'assumed consent' where an email or app notification is sent which indicates to the user that by continuing to use the service, they are agreeing to the data sharing. Third, there is 'opt-out consent' that provides an option to specifically disallow the data to be shared. Fourth, 'opt-in consent', where the data is not shared by default and requires explicit permission.

3.3.5 Privacy Mechanisms

Our survey includes questions on how acceptability is influenced by privacy mechanisms. The five privacy mechanisms we included are local differential privacy (LDP), central differential privacy (CDP), data anonymization, data aggregation, and encryption [155, 237,

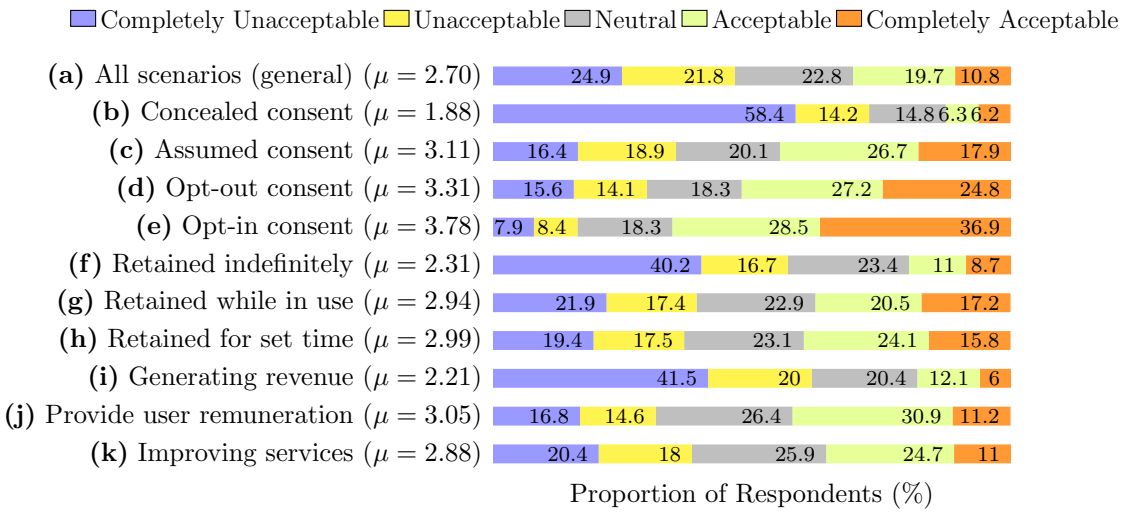


Figure 3.2: The acceptability distribution of multiparty data sharing across all scenarios for each variable. Acceptability is measured on a five-point semantic difference scale and each segment corresponds to the proportion of respondents who select that level of acceptability ($N = 916$).

245]. Respondents each received one of the five privacy mechanisms and rated the acceptability of the data sharing scenario, if it were to include that privacy mechanism. To validate that respondents understood the mechanisms, our research team manually generated informal descriptions of the mechanisms, and the survey asks respondents to match their privacy mechanism to the most accurate description. This unfortunately suggested respondents had low comprehension of the privacy mechanisms provided to them. Thus, we exclude privacy mechanism related results.

3.3.6 Demographics

We report an overview of demographics rounded to the nearest percent. All survey respondents are located in the United States. Of the total $N = 916$ participants, when asked to specify their gender, 47% specified man, 50% specified woman, 1% specified non-binary, 2% preferred not to say, and less than one percent chose to self-describe. Respondents specified an age range with 17% of respondents selecting 18-24, 22% 25-34, 15% 35-44, 21% 45-54, 22% 55-64, and 3% prefer not to say. In terms of employment, 70% reported the industry of their current form of employment, 18% reported being unemployed, 5% as student, and 6% responded with prefer not to say. The industries reported by those that were employed were diverse with the most frequent industry being education at 10%. A

slight majority of participants reported completing a degree at 59% (bachelor, graduate, or associate). The remainder of participants education can be broken down as 23% with some college but no degree, 14% completed high school, 3% less than high school, and 1% prefer not to say.

3.3.7 Limitations

We recognize that our scenarios are not all encompassing of multiparty data sharing. We have included varying companies, data types, and structures such that it may guide the focus of future work. The companies we selected for this study include a focus on health companies and health data. This focus may have influenced respondents in hard to predict ways based on respondents presumptions about how health data is regulated as well as their willingness to share such data. Further, we use a semantic differential acceptability scale, but acknowledge that such scales could still result in bias over the duration of the questions presented. Responses were gathered while the COVID-19 pandemic was ongoing [43]. We cannot know how this may have affected respondents' answers, but it may have contributed to the higher unemployment percentage.

We further note that our participants, from across the United States, are WEIRD (Western, educated, industrialized, rich and democratic) [198]. We do not presume to make global assertions from our study but instead show that even within this group there is a diverse set of expectations and preferences not currently supported by technology nor required by regulations. Our scenarios are based on examples located in North America, where our respondents live. This is critical as different regions, even within WEIRD participant pools, have different existing laws and expectations. For example, EU citizens already have different protections than non-EU citizens. Finally, we acknowledge the potential for bias towards perceived socially desirable behaviour [186]. We attempt to mitigate this bias by using the more neutral term 'acceptable'. We ensure there are no mentions of privacy until the end of the survey, and we give participants the opportunity to provide their own views in free-form text.

3.4 Results

We first present respondents' overall perceptions of multiparty data sharing and related user controls. Second, we examine the differences in acceptability between and within each sharing type. This is followed by our analysis of demographic based variations in

perceptions. Finally, we present an exploration of respondents' free-form responses. Recall, the labels for the variables and collaboration types are found in Table 3.5.

The results we present highlight our statistically significant findings. For interpretability, we report mean values for acceptability in this section. When we refer to statistically significant differences, we are not referring to these means, but include them as the statistical mean ranks are less interpretable.

Note that although we asked respondents questions with respect to how privacy mechanisms could impact acceptability, unfortunately respondents' comprehension of the privacy mechanisms definitions was low (based on our validation definitions), however it is included for reference at the end of this section.

All statistical results presented use a significance level of 0.05. We use non-parametric statistical tests as our data is not normally distributed. This leaves the potential for incorrectly finding a difference insignificant. However, it decreases the risk of incorrectly saying a difference is significant. Additionally, when presenting the results of multiple comparison procedures, we report the p-value adjusted using the Bonferroni correction to account for the increased chance of false positive results due to multiple comparisons.

3.4.1 Overall Perceptions

We begin by determining a base understanding of how acceptable respondents find multi-party data sharing and our defined variables, regardless of the type of collaboration they received. The acceptability of the data sharing scenario in 'general' (a), is completely unacceptable or somewhat unacceptable to 45% of respondents. Without additional details about the collaboration, participants respond slightly more towards the unacceptable end of the scale, but almost 30% of respondents do find it to be at least somewhat acceptable. The distributions of how acceptable respondents found each variable are shown in Figure 3.2.

We perform a Friedman's two-way analysis of variance by ranks for each of the distributions of acceptability: within informed consent groups, within data retention groups, and within purpose groups. For all variables within groups $N = 916$. Results show that the distributions of acceptability is not the same for: consent groups (b), (c), (d), (e) with test stat = 899.29 p <0.001, retention groups (f), (g), (h) with test stat = 255.08 p <0.001, and purpose groups (i), (j), (k) with test stat = 435.79 p <0.001.

We perform Dunn's multiple comparison procedure to identify which variables within a group differ and in what direction, for example within data retention, how do variables (f),

(g), and (h) differ (see Table 3.4). The difference in mean rank (e.g., the mean rank of Var 1 subtract the mean rank of Var 2) shows the direction of the difference in acceptability of the pair. All pairs of variables have significantly different distributions of acceptability except for the (g), (h) variable pair from within data retention.

Within Informed Consent. All user control variables for consent, (b) through (e), have statistically significant differences in terms of acceptability. Overall, in terms of notification and consent, participants find data sharing more acceptable when they are explicitly informed or have more control over whether their data was used. 'Concealed consent', when they receive no formal notification, is overwhelmingly unacceptable to 73% of respondents ((b), $\mu = 1.88$). Unacceptability is substantially reduced when users are notified in any manner, regardless of control (e.g., even if opt-in or opt-out options are not available). 'Opt-out consent' ((d), $\mu = 3.31$), where users can toggle a setting to indicate they do not want their data shared, skews slightly more towards the acceptable end of the scale than the 'assumed consent' case ((c), $\mu = 3.11$). 'Opt-in consent' achieves the highest acceptability ((e), $\mu = 3.78$) within the consent/notification grouping with approximately 58% of respondents finding it at least somewhat acceptable.

Within Data Retention. We investigate respondents' perceptions with respect to data retention, (f) through (h), and find significant differences in their acceptability. Respondents find 'retaining data indefinitely' ((f), $\mu = 2.31$) to be less acceptable than retaining the data until the company is 'finished using it' ((g), $\mu = 2.94$) and less acceptable than retaining the data for a 'specified time' limit ((h), $\mu = 2.99$). There is no significant difference in the distributions of how acceptable respondents find data between retention for a 'set period of time' and 'as long as the company uses it'. However, in practice there could be no real difference in how long the data is retained between indefinite retention and retaining the data as long as the company is using it. This result highlights the risk of influencing users consent based on phrasing; something not currently strictly defined across regulations on data sharing.

Within Purpose. In terms of purpose of use, (i) through (k), there are statistically significant differences in how acceptable respondents find each purpose. Respondents' overall perceptions are summarized as follows. It is least acceptable when the company (or companies) uses the data to generate revenue ((i), $\mu = 2.21$). Respondents find it somewhat more acceptable when there is an explicit tangible or perceived benefit to the user, such as a monetary reward ((j), $\mu = 3.05$) or improved service ((k), $\mu = 2.88$).

Var 1, Var 2	Difference in Mean Rank	Std. Test Statistic	p
	Informed Consent		
(a), (b)	-0.85	-14.098	<0.001
(a), (c)	-1.07	-17.663	<0.001
(a), (d)	-1.44	-23.798	<0.001
(b), (c)	-0.22	-3.565	0.002
(b), (d)	-0.59	-9.700	<0.001
(c), (d)	-0.37	-6.135	<0.001
	Data Retention		
(f), (g)	-0.46	-9.778	<0.001
(f), (h)	-0.51	-10.864	<0.001
(g), (h)	-0.05	-1.086	0.832
	Purpose		
(i), (j)	-0.71	-15.186	<0.001
(i), (k)	-0.55	11.764	<0.001
(k), (j)	-0.16	-3.423	0.002

Table 3.4: Dunn's multiple comparison test results for the distribution of acceptability compared pairwise between variables within informed consent, data retention, and purpose groups. All p-values are adjusted for multiple comparisons (6 comparisons for the consent group, 3 for each of the data retention and purpose groups).

3.4.2 Nature of Collaboration

Recall the five types of collaboration defined in Section 3.3.1 and shown in Figure 3.1. First, we examine between group differences, that is, the differences in acceptability between different collaboration types. Second, we present within group differences, more specifically, the difference in acceptability between the scenarios that comprise a collaboration type.

Between Collaboration Types.

We compare our five types of multiparty data sharing to investigate whether some sharing types are more acceptable to respondents. The different average acceptability scores across types of collaborations for variables (a) to (k) are shown in Figure 3.3. To determine

Variable	**Label**
All scenarios (general)	(a)
Concealed consent	(b)
Assumed consent	(c)
Opt-out consent	(d)
Opt-in consent	(e)
Retained indefinitely	(f)
Retained while in use	(g)
Retained for set time	(h)
Generating revenue	(i)
Provide user remuneration	(j)
Improving services	(k)
Collaboration Type	**Label**
Validation	(V)
Two-way Two-Party Exchange	(1)
One-way Two-Party Exchange	(2)
Many-to-One Exchange	(3)
Acquisition	(4)
Merger then Acquisition	(5)

Table 3.5: Reference table for labels corresponding to usage controls and collaboration types.

which types of collaboration are more or less acceptable we perform a subsequent pairwise analysis.

With respect to acceptability in 'general' (a), the different collaboration types, (1) through (5), are statistically significantly different. Both 'acquisition' ((4), $\mu = 2.96$) and 'merger then acquisition' ((5), $\mu = 2.93$) are more acceptable than a 'one-way two-party exchange' ((2), $\mu = 2.51$) and 'many-to-one exchange' ((3), $\mu = 2.34$). A possible attribution to the greater acceptability for mergers and mergers then acquisition rather than exchanges could be the indirectness by which data is acquired. Unlike in the specific exchange scenarios ('one-way two-party' and 'many-to-one') where data can be seen as a commodity, within the merger-acquisition scenarios nobody is explicitly seen as 'selling' users' data. Additionally, in the case of mergers and acquisitions, the company acquiring the data may be seen as the new shepherd of the data, continuing to provide the user with the services that led them to originally use the acquired companies' services.

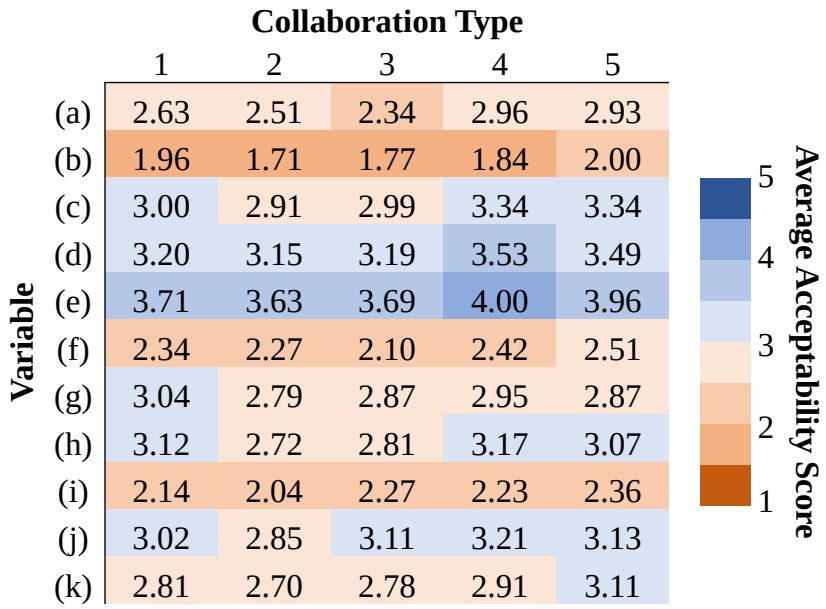


Figure 3.3: Average acceptability of variables for each collaboration type. The labels for collaboration types and variables correspond to those shown in Table 3.5.

User Controls Between Collaboration Types. We further compare between collaboration types for each of the user control mechanisms. We continue to observe statistically significant differences between mergers and acquisitions compared to the other exchange types. Specifically, 'improving services' (k) is more acceptable for a 'merger then acquisition' ((5), $\mu = 3.11$) than a 'one-way two-party exchange' ((2), $\mu = 2.70$). 'Assumed consent' (c) is more acceptable for both merger collaboration types ('acquisition' ((4), $\mu = 3.34$) and 'merger then acquisition' ((5), $\mu = 3.34$)) than for a'one-way two-party exchange' ((2), $\mu = 2.91$). Finally, 'retained for a set time' (h) is more acceptable for an 'acquisition' ((4), $\mu = 3.17$) than a 'one-way two-party exchange' ((2), $\mu = 2.72$) The difference in acceptability between types for data retention and purpose could again be potentially attributed to the indirectness by which data is acquired in mergers and acquisitions.

There are no notable differences between collaboration types for 'concealed consent', 'generating revenue', 'retained indefinitely', and 'retained while in use'. This unchanging negative perception is likely because these attributes are considered more uniformly unacceptable. These results demonstrate another avenue where users would benefit from

transparency in terms of the purpose and other contextual information, to make an informed decision of whether to consent, when companies are merged or acquired.

Within Collaboration Types.

Each collaboration type consists of two possible scenarios. We compare the scenarios within each collaboration type to one another to identify differences that exist depending on the sending and receiving companies as well as who the respondent is a user of. In our analysis we do not consider the order that the companies are introduced as a factor. This exclusion is based on our validation test for collaboration Type V; which found no statistically significant differences between the response distributions whether a health or technology company is introduced first, across variables (a) through (k).

We summarize the remainder of our results within collaboration types by their common themes. Overall, the within collaboration types analysis suggests that the inclusion of a health company negatively influences users' perceptions of the multiparty data sharing.

Collaboration over Commodification for Health Data. We find an interesting result within the 'one-way two-party exchange', an exchange type where the key distinction between scenarios is a tech company giving away user data (Scenario E) versus a health company giving away user data (Scenario F). We identified statistically significant differences across seven of the eleven measured variables. The four non-significantly different variables are 'concealed consent', 'assumed consent', 'opt-out consent', and 'retained indefinitely'. For the seven variables that do have significant differences, they are all more acceptable for Scenario E when compared to Scenario F. In Scenario E, respondents are framed as a user of a technology company which is providing its data to a health company. Whereas, in Scenario F respondents are framed as a user of a health company which is providing its data to a technology company. In both Scenario E and F, respondents are a user of the company giving away data.

This suggests the difference in acceptability could be attributed to the commoditization of health data being more objectionable than in the case of tech data. While respondents may be used to, or even have come to expect to have their data treated as a commodity by technology companies (Scenario E), the same may not be true for health companies. To further this idea, we look within 'two-way two-party exchanges' (Scenarios C and D), wherein the health company shares its data but also receives data in return. Respondents seem to interpret this reciprocity as providing some benefit to them, as opposed to being a 'sale'. When this reciprocity is absent in Scenario F, we see lower acceptability overall,

possibly due to this commodification of health data which has an expectation to be the most protected data.

Health Companies Complicate Data Sharing. Health companies being involved negatively impact user perceptions of multiparty data sharing even when the health company is only receiving data. This is shown, first within 'many-to-one exchange', wherein a number of companies are sharing data with either a tech company (Scenario G) or a health company (Scenario H). We found a significant difference in acceptability of 'assumed consent'. Respondents who received the scenario where a technology company acquired the data (Scenario G, $\mu = 3.25$), found 'assumed consent' to be more acceptable than when a health company received the data (Scenario H, $\mu = 2.76$). This result implies that users were not as satisfied with simply being informed of data sharing, when it is shared with a health company, in contrast with a technology company. As both scenarios involve sharing financial data, we can hypothesise that users do not want their financial records to influence any future medical diagnoses. Users may be concerned for discrimination while receiving medical treatment or processing insurance, if a health company obtained their financial records.

The negative impacts of health company in data sharing is also shown within the 'acquisition' collaboration type. Scenarios within 'acquisition' involve a start-up that tracks user data on diet, fitness, and social habits being acquired by either a technology company (Scenario I) or a health company (Scenario J). Respondents found 'opt-in consent', the "strictest" consent option of the ones we tested, to be more acceptable when a technology company (Scenario I, $\mu = 4.24$), rather than a health company (Scenario J, $\mu = 3.77$) acquired a startup. We expect that respondents are more comfortable with their fitness habits influencing technology products, like in Scenario I, rather than having the potential to influence their medical treatment or insurance as in Scenario J.

As a final note on the inclusion of health companies and how they may influence respondents, we note that health data has certain laws surrounding it that respondents may believe will protect them. Further, respondents concerns with data transferring to or from a health company may also be attributed to respondents being unsure as to the purpose. From our free-form responses we know that the purpose of use for the data was a frequent condition for acceptability.

3.5 Privacy Mechanism Comprehension

Respondents predominantly fail the comprehension check as to whether they understand their privacy mechanism. Only 37% of total respondents correctly identified the corresponding "layperson" description of the privacy mechanism they received. Data aggregation was the most correctly identified with 64% correctness. Respondents had the most difficulty comprehending LDP and CDP. As LDP and CDP are essentially modifications to aggregation when described less formally, it is not surprising that they were frequently thought to correspond to the aggregation description. Privacy mechanism comprehension results are shown in Figure 3.4.

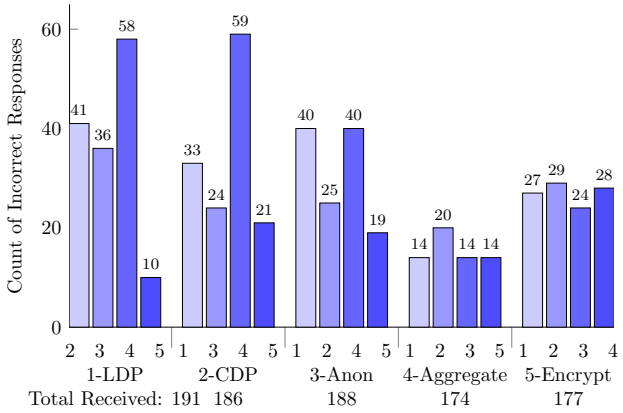


Figure 3.4: The counts of privacy mechanism received versus incorrectly guessed. Respondents receive the definition of a privacy mechanism and attempt to identify the layperson description that corresponds to that same privacy mechanism. For example, of the 191 respondents that received LDP (privacy mechanism 1), 41 incorrectly guessed they received CDP (privacy mechanism 2).

3.5.1 Demographic Variations

We evaluate responses across all scenarios for differences based upon demographic groupings.

Gender acceptability variations. For gender, we performed a Mann-Whitney U Test with two groups comprised of 432 men and 455 women compared for each of the variables (a through k). We found a significant result for 'concealed consent' (variable (b)). We can conclude that men found their consent not being explicitly granted, to be significantly

more acceptable than women did ($p = 0.008$, std. test statistic= -2.647). The difference in mean rank between men and women for 'concealed consent' was 40.45.

Age acceptability variations. To examine how age group influences acceptability for each of the variables, (a) through (k), we performed a Kruskal-Wallis test comparing the five age groups {18-24 ($N = 154$), 25-34 ($N = 201$), 35-44 ($N = 140$), 45-54 ($N = 197$), 55-64 ($N = 201$)}. We find that between age groups, the acceptability distribution of (a) $p = 0.006$, (b) $p < 0.001$, (f) $p < 0.001$, (g) $p = 0.019$, (h) $p = 0.012$, (i) $p < 0.001$, (j) $p = 0.018$, and (k) $p < 0.001$ is not the same.

Demographic variations overview. For demographic differences due to gender, we compare men versus women as we did not have enough respondents representing other genders, leaving us with $N = 887$. We compare the two groups comprised of 432 men and 455 women across the variables (a through k). From our data we identify a significant difference in (b) 'concealed consent'. We can conclude that men ($\mu = 1.98$) found their consent not being explicitly granted, to be significantly more acceptable than women ($\mu = 1.76$) did.

To examine demographic variations due to age, we compare five age groups. From our data we identify a significant difference due to age group across all variables except for 'assumed consent', 'opt-out consent', and 'opt-in consent'. Across the variables 'concealed consent', 'retained indefinitely', 'generating revenue', and 'improving services', respondents aged 55-64 find each variable to be significantly less acceptable than their otherwise aged counterparts (18-24, 25-34, 45-54). Respondents aged 35-44 find the 'general scenario', 'retained for set time', and 'generating revenue' less acceptable than respondents aged 45-54. Additionally, those aged 35-44 find 'improving services' less acceptable than respondents aged 18-24. While older people, such as our respondents aged 55-64, could be generally expected to have more conservative views, we do not know why the middle age group, respondents aged 35-44, have a similar lower level of acceptability across variables. These results demonstrate that different demographics have different desires. User controls need to have sufficient individualization to support these differences.

3.5.2 Free-from General Perceptions

We analyze 789 non-empty free-form responses to the question 'In general, what are your thoughts on companies sharing data with other companies'. We exclude 62 responses that

are either not interpretable or indicated a desire not to respond. For example, exclusions include: single or random character responses (e.g., 'a', 'alskj'), 'N/A', and 'no'.

Responses were coded in terms of their positive or negative perceptions of the practice of sharing data. Positive or negative responses can have a conditional component that indicates what improves or worsens their perceptions. The codes for the free-form responses perceptions were developed through discussion and definition by two members of the research team based on a sampling of the response set and the predefined 'positive' and 'negative' codes. A 'resigned' and a 'neutral' code were added after initial sampling to more accurately describe all responses. This methodology follows the process of Oates et al.'s [160] analysis and Miles et al.'s Qualitative Data Analysis: A Methods Sourcebook [151].

The final codebook used to code the free-form responses is *neg* for unconditionally negative, *neg. Cond.* for overall negative response but permitted cases, *neutral* for neither positive nor negative, *resigned* for negative but accepted, *pos. Cond.* for overall positive but forbidden cases, and *pos.* for unconditionally positive.

Two members of the research team each independently applied the perceptions codebook to the response set. The coded responses were reviewed for agreement by the two team members. The process for handling a disagreement in coding was for both coders to check their responses. If the difference could be attributed to having mislabelled the code, a correction would be made. The coders would come back together and check the new agreement. If disagreement persisted, it went to a tie-breaker coder. We include an overview of common themes that were indicated as influencing perceptions or requirements in 'conditional' responses. The final code counts are summarized in Table 3.6.

Polarity of Base Perceptions

Of the total (789) responses coded for positive and negative perceptions, 32 required a third coder to break the disagreement. The original two independent coders agreed on the codes for 757 responses, or 96% of responses after checking for errors. One of the 32 responses shared with the third coder was coded differently by all three coders and the consensus was to remove it due to the ambiguity.

Unconditionally negative responses formed the largest group of responses and included a breadth of subjects relating to purpose, laws and regulations, distrust, and risks. Objections include users' data being used for generating revenue for the company or for marketing purposes.

	Frequency by Collaboration Type					
Code	**All**	**1**	**2**	**3**	**4**	**5**
Neg.	305	46	45	54	49	48
Neg. Cond.	107	19	15	12	17	22
Neutral	66	12	7	10	15	16
Resigned	32	6	4	5	7	10
Pos. Cond.	165	26	25	26	33	37
Pos.	51	5	10	4	13	8

Table 3.6: Frequency given Nature of Collaboration. Columns correspond to: 1. One-way two-party exchange, 2. Two-way two-party exchange, 3. Many-to-one exchange, 4. Acquisition, and 5. Merger then acquisition.

> P58310: I think companies after having acquired data as an asset has one intention and it's making money through exploitation"

> P78909: "These companies are reprehensible! I will not consent to my data being shared for marketing purposes"

Other negative responses report distaste for being coerced into agreeing to data sharing in order to access services. Respondents consider such requirements an uneven trade given the risks associated with a breach that exist whenever data is collected, saved, and shared.

> P20322: "I'm not happy about it because if you do agree you can't choose who it will be shared with. If you don't agree, you can't use the service"

> P53560: "I hate it. Cookies and data thieves. Opting out often renders the website unaccessible- so it's coercion/entrapment. Data breaches wouldn't really happen if data wasn't retained"

Other possible risks of such sharing, according to respondents, include malicious outsiders and malicious companies. Respondents express concern with targeted manipulation by a company, such as advertising, using the shared data. Concern with breaches or leaks also includes concern for data leaking out in ways users do not expect.

> P69036: "While there are clear and logical reasons for utilizing and selling this data it does have potential for targeted manipulation"

> P36717: "no. It makes me feel like my personal information is keep leaking out. i feel more vulnerable"

The responses coded as 'resigned' essentially express that respondents know such sharing occurs, do not necessarily like it, but accept it as reality. Respondents also express a need for law or regulations, a belief that such events are likely more or less frequent than they know, a feeling of futility, and the implied agreement to such things when using apps. One participant's response encompasses each of the above themes.

> P07944: "It's a gray area: users make and agreement with companies for information use based upon the scope and identity/reputation of a company. What happens with an individual's information in the event of a the business/organization being sold. Legally speaking, the matter is an open and shut case. However, a user may not want to have the same information use agreement with the new company...and their rights to having a say in how their information is being used are clearly being violated by the new company which technically owns the rights to the information they have purchased since the company never negotiated terms with users and can use that information according to the company's desire and purposes. It's legal; but it sucks"

The neutral responses include two main types. First, some respondents directly say they are neutral or do not care about such sharing. Second, some respondents express some potential limitations on such sharing, but that they still did not have strong feelings about it either way.

> P79659: "I don't have definite objections to companies sharing data with other companies"

> P60109: "It depends on what it's used for and must have complete consent from an individual that isn't forced"

Few of the unconditionally positive responses say more than a one to three word answer. For example, 'good', 'epic', and 'sounds great' are common. The positive responses beyond sharing a generic response include some benefit to the individual or to the company. Benefits include personalizing advertising, ad opportunities, and new developments. While distaste for data being used for advertising was found in many negative responses, such as the earlier examples, this distaste was not universal.

P14505: "I think that it is acceptable because they need to use this data for advertising opportunities"

P98147: "Data sharing encourages more connection and collaboration between researchers, which can result in important new findings within the field. In a time of reduced monetary investment for science and research, data sharing is more efficient because it allows researchers to share resources"

Conditionals and User Control

In this section we focus on the responses coded as conditional. We highlight requirements users report as necessary for the scenarios to be acceptable. Specifically, we review 'positive conditional', 'negative conditional', and 'neutral' coded responses with respect to their conditionals. We include 'neutral' as our code definition of 'neither positive nor negative', does not prevent conditions from being specified in the response. Whether respondents viewed the scenario positively or negatively, they expressed similar themes.

Consent. The importance of consent and transparency is prominent in both positive and negative conditionals, with an emphasis on informed consent. Respondents express a need for easily accessible opt-out options and that consent (to data sharing) should not be a requirement for using a service.

P66884: "It's inappropriate unless the user consents explicitly and should never be a requirement for use"

P10652: "I do not think it is acceptable unless they have the users permission. Or an option to cancel information sharing. If the user has a choice and is OK with it then I believe it's fine"

P19193: "If they make people aware (in BIG print, not small, easy-to-miss print) then it's fine"

When expressing the importance of users' consent, some respondents highlight that data sharing should not be taken lightly. There are risks that can be associated with data being provided to other entities that cannot be properly evaluated without details as to where the data is going, what the data is, and why it is being shared. Receiving user consent requires full transparency with respect to each of those attributes.

> P91741: "It should not be shared unless the individual gives authority to do so. It is private information that should not be shared on a whim"

> P09262: "I don't think companies should share customer's personal information unless specific consent is received from the customer to where/what the information is shared to, as well as why"

Furthermore, consent can be withdrawn and cannot be assumed to be transferable between entities, even in the case of a company being purchased.

> P41281: "Information collected, with the users permission, should never be shared with another company or assumed to be the property of said company if they merge with another company. This would be true regardless of whether the original company remains in the same business, or moves into a different service."

However, some respondents highlight that sufficient transparency can be advantageous to companies building goodwill after mergers or acquisitions.

> P48036: "...The company can email its acquired users and them that they bought out the nicestartup and they want to use the data in order to improve their services and then list their services so people can decide for themselves. You'll be surprised how many people will agree to continue, there's no need to hide, lie, or manipulate anything. Just be honest! You'll earn respect and loyalty as well"

Data Type and Processing. Respondents indicate preferences for the kind of data and how the data is processed.

> P31222: "I do not like the idea of any personal, individual information being shared with other companies, either for free or for a price, but if a study is performed on that data and then the study results are shared I completely think that is okay"

The type of data that is acceptable or unacceptable is not universal. Respondents mention opposition to medical or health data generally, although there is some acknowledgment of

possible exceptions. While personally identifiable information (PII) is generally expressed as inappropriate to share, what counts as PII is less universal. Some respondents consider buying habits to be fine while others highlight the private nature of such financial transactions [53].

> P45732: "I don't mind sharing information as long as it's not financial"

> P71169: "I have a problem with this when it's sensitive personal information such as health information. I don't have as much of a problem with this when it's something less sensitive, such as my buying habits"

Purpose. The acceptability of different data sharing purposes, at least as far as the free-form responses are concerned, is highly individualized to what each respondent considers beneficial or detrimental. Some respondents find advertising acceptable while others do not. Sharing data to improve services or scientific investigations are spoken of positively while selling users' data for monetary gain is aggressively opposed.

> P24797: "It depends upon the purpose (my benefit or detriment), the data security to ensure the original personally identifiable data is secure or destroyed and the trust based on the history of how the company previously handled data"

Health. Health data is the most controversial type of data sharing, and a number of respondents express concern for whether legitimate sharing purposes exist. Many respondents that mention health data do so with intense negativity and concerns over the relevant ethics and legality of the exchange or purchase of health-related data.

> P94865: "Repugnant, especially in light of for-profit health systems attempting to maximize profitability from patient interactions"

> P72271: "There are stringent rules about sharing data under HIPAA in the US and this clearly violates it, along with potentially exposing PII"

> P77878: "worried that data will be mined for insurance companies so they can eliminate or remove costly illnesses"

Even within the topic of health data, some respondents reflect upon the potential for acceptable data sharing settings. Privacy protections are key to improving the acceptability of health data sharing. Protections could include regulations, privacy mechanisms, and greater transparency.

> P20986: "It depends. I think it can be beneficial under certain circumstances, but I would be hesitant having any healthcare data shared outside my practitioners. However, I recognize how it can improve goods/services, but there has to be a lot of protection in place anytime data is shared"

> P44838: "I believe for health records it should be acceptable for continuance of care but not for advertising or making money"

3.6 Discussion

Disambiguate Third Parties. Privacy policies that give companies unrestricted ability to share data with 'third-parties' and 'partners' do not encapsulate the details that influence users' preferences. Our results show users care about who data is being shared with, what is being shared, and the structure of the collaboration. In terms of 'who', health companies sharing data is less acceptable than technology companies. In terms of 'what', it is more acceptable to share fitness data with a technology than a health company. Structurally, reciprocity improves acceptability over one-way 'sale' type transactions. Transparency with respect to the nature of any collaboration is required to support the preferences our respondents expressed. Thus, regulations, such as CCPA, need to have detailed requirements for companies to clearly outline the properties we identify for data sharing.

Explicit over Implicit Consent. Implied consent is inferred based on a person's actions or circumstances. When companies make consent conditional for the use of their service, the use of the service is taken as consent. In contrast to implied consent, explicit consent is unmistakably provided by the user, possibly in writing. It is specific, can be rescinded, and is non-transferable. Informed consent requires users to have an understanding of the implications and extent of what their agreement applies to when using an app or tool. Respondents in our study expressed a clear preference for explicit consent that requires them to opt-in over implied consent (e.g., 'concealed consent' or 'assumed consent'). Respondents' preference for, and emphasis on consent and transparency, held for both statistical analysis and free-form responses.

Reduce Ambiguities to Communicate Privacy. Although user controls affect the acceptability of collaborations, the effect does not always correspond to the impact on privacy in practice. For example, retaining data 'while in use' and 'indefinitely' may have no practical difference. Despite this, respondents found it more acceptable for companies to retain data 'while they are using it'. Companies could abuse such misunderstandings by making something seem more private in practice than it actually is.

Similarly, each of the five privacy mechanisms we included have a different effect on privacy in practice. Respondents to our study had difficulties understanding our descriptions of the privacy mechanisms. Unless users can distinguish between accessible descriptions, they will not be making informed decisions. Therefore, when companies use privacy mechanisms, they should be compelled by law to ensure it is either easy to understand or that users are not required to understand the privacy mechanism used to successfully make an informed choice. Going forward, researchers and policy makers must focus on conveying the significance of different privacy implications and changing the information provided to users such that it is clear and concise and not perceived as minor details.

Consent, Notice, and Choice. While the participants in our study expressed a strong valuation in being able to give informed consent, it is important to contextualize this result with respect to what this actually means for individuals. As has been shown in the past with mobile app permissions and more recently with cookie banners, the use of persuasive design (also sometimes referred to as dark patterns) by companies to manipulate users to agree to certain terms is a risk [88, 127]. Thus, there are two key aspects related to this issue that remains to be explored. First, what does it mean to be able to consent to data-sharing practices. What levels of control do users want and need? Can the security community implement different baselines that better reflect what users want and need? For example, it may be the case that we could develop a new series of data-safety standards that are implemented across a region. These could be similar to other safety practices such as the food and drug regulations and automobile safety. However, such standards require the research community to know what risks exist for certain practices if we are to similarly create warning labels such as those that exist for lead, alcohol, tobacco, and others. While efforts to make "nutrition labels" for privacy have previously been attempted [116], we still have not found a solution for this space. The second aspect that needs to be explored further is what ways of communicating to users about data-sharing are both informative and non-manipulative. Without knowing how to do this, it will not just be difficult to inform users but it will also be difficult to protect users from companies that employ manipulation tactics. Going forward, researchers, developers, and policy makers will all have to work together to address and advance these two key aspects.

3.7 Conclusion

We presented the results of our survey on user perceptions of multiparty data sharing. Our results indicate that the type of data sharing collaboration affects acceptability as do the available user controls. Based on these results, we recommend that regulations for data sharing do not solely rely on past work that focused on only one company receiving data from another (whether for advertising or other purposes). We hope the recommendations we have made help other privacy researchers and regulators mitigate the inequity imposed on users by data commodification.

Chapter 4

Practicalities: Theory to Practice

This chapter is adapted from work that previously appeared as "Mind the Gap: Ceremonies for Applied Secret Sharing" at the 2020 Privacy Enhancing Technologies Symposium [109]. In this chapter we highlight how even well established privacy technologies, such as secret sharing have practical privacy failures when not analyzed with respect to human-factors. Through this chapter we will show that incorporating human-factor based analysis, such as ceremony analysis, allows us to better develop technical protocols that meet the privacy needs of the application setting; where otherwise there would be remaining vulnerabilities. With respect to privacy in machine learning this is a critical result for designs going forward, many of which employ conventional cryptography techniques such as secret sharing. In order to develop strong privacy-preserving machine learning; a better understanding of the deployment settings in terms of human-aspects can increase the success, in terms of privacy protections, for such applications.

4.1 Introduction

The security properties that theoretical secret sharing purports to provide are particularly meaningful for high-risk users such as journalists, as demonstrated by the security-critical effort required for the investigation and reporting of the Panama Papers [143]. However, while the security of theoretical secret sharing is well documented in academic research, in practice, the security guarantees are more complicated.

The descriptions in the literature of secret sharing schemes, which we additionally refer to as threshold schemes, often lack sufficient evidence of the security of real-world deployments of the schemes. This shortcoming is due to the descriptions leaving a large number

of assumptions and decisions to the participants, as these are considered to be outside of the protocol. Just as the design of the highly successful TLS protocol accounts for more real-world practicalities than the underlying Diffie-Hellman key exchange protocol, the practical use of threshold schemes, as we demonstrate, is no different. For example, although Shamir secret sharing (see Section 4.2) is information theoretically secure, ultimately the shares must be communicated to participants through a channel, which in most cases will rely on symmetric or asymmetric encryption, and therefore rely on computational assumptions. Furthermore, unlike cryptographic protocols such as Diffie-Hellman, threshold schemes require significant user involvement and decisions at nearly every stage of the protocol. Consequently, analyzing the security of threshold schemes requires assessing both the protocol and the actions and decisions required of users.

Ellison [64] introduced the concept of a *ceremony* in security analysis, which requires the inclusion of both the cryptographic *protocol* as well as any possible *user actions or decisions* in the security analysis. Surprisingly, the state of research literature for threshold schemes does not include a complete, end-to-end, formal definition and assessment of the security of the ceremony of threshold schemes. Without such definitions, deployments of threshold schemes lack the necessary structure required for formal analysis as their flexibility in terms of applications is broad.

Without strict boundaries for a specific threat model and use case, it is impossible to provide both a generalized framework and a formal analysis for secret sharing ceremonies. This work provides a structure, in the form of a framework, that can be used for a given threshold scheme to define and analyze its particular ceremony, by structuring the ceremony as a series of stages and steps as is necessary to assess the ceremony's end-to-end security. Although an unbounded number of possible user interactions exist, our framework can be used to guide the definition and formalization of the ceremony. We identify ceremony-related issues, such as requiring the dealer to delete sensitive material and the requirement for users to authenticate one another. The ceremony for an application, defined in terms of our framework, accounts for the specific goals and adversaries of the ceremony and therefore provides the needed structure that must precede efforts to formally analyze a specific ceremony.

Contributions. We provide a framework to facilitate the process of defining an accurate ceremony for a given threshold scheme; we then use this framework to assess the security of several threshold schemes, and define a lightweight set of improvements that are useful to threshold schemes based on Shamir secret sharing. Our framework is useful for comparing different existing secret sharing schemes; however, what it primarily provides is a structure for defining threshold scheme ceremonies with the necessary details to perform a more accurate security analysis that accounts for the setting in which the threshold scheme is

used. Overall, our contributions include:

- a demonstration of the variability in the ceremony of threshold schemes and how this variability can lead to gaps in the security properties achieved by the threshold scheme;
- formalizations of the adversaries and of several use cases of threshold schemes used in practice;
- a framework to facilitate security analyses of threshold schemes used in real-world settings;
- exemplar applications of our ceremony framework via security analyses of three threshold scheme case studies; and
- techniques to close security gaps uncovered in our above analysis and an implementation of these improvements in Rust.

Organization. This chapter is organized as follows. Background, motivation, and related work are Sections 4.2, 4.3, and 4.4 respectively. Section 4.5 is our framework for our analysis. Formalized ceremonies for two modes of operation for threshold schemes are in Sections 4.6 and 4.7. Section 4.8 summarizes our analysis for several threshold schemes, Section 4.9 is our improved ceremony and implementation and our conclusion is Section 4.10.

4.2 Threshold Schemes

We summarize the notation used throughout our analysis in Table 4.1, including both notation standard to the literature as well as new notation we introduce in later sections for the purpose of our analysis. Notably, we denote the secret information that is protected by the threshold scheme as $\mathcal{F}$, while the secret input into the threshold scheme is s. This differentiation will become important when we define modes of operation where $\mathcal{F}$ can either be equal to s, or distinct, as defined in Section 4.5.2.

In general, cryptographic secret sharing schemes enable a group of n participants, possessing a secret s, to divide s into n shares. Before creating the n shares, a threshold value t is chosen such that a collection of t shares must be used to learn the value of s. A

n	number of participants
t	threshold
s	secret recovered by a (t, n)-threshold scheme
$\mathcal{S}$	secret space of a (t, n)-threshold scheme
D	dealer
D_P	dealer that later becomes a participant
$\mathcal{F}$	sensitive information requiring protection
Base	$s = \mathcal{F}$
Ext	$s \neq \mathcal{F}$
P_r	participant performing a recovery of s
$\mathcal{U}$	participant performing an update
$\vec{C}$	commitment used to validate any share

Table 4.1: Parameters and additional notation used within our analysis

(t, n)-*threshold scheme* is a secret sharing scheme where n and t are positive integers such that $t \leq n$, n representing the number of participants, and t the desired threshold.

In a (t, n)-*threshold scheme* we designate a *dealer* D as the entity that selects the *secret* s and generates the n shares such that each of the n participants in the scheme receives a share that preserves the following properties:

Reconstruction: any size-t subset of the n participants can compute the secret given their t shares, and

Secrecy: no subset of the n participants consisting of $t - 1$ or fewer participants is able to gain any knowledge of the secret given their combined shares.

In a conventional (t, n)-*threshold scheme*, the set of n participants does not contain the dealer D. However, in our analysis we work in the setting where the dealer, now labeled a *participant dealer* D_P, may continue to be involved in the scheme as a participant, as typically occurs in real-world practical settings.

While some variants of threshold schemes, such as threshold signature schemes [84], allow participants to use their shares of s individually and to perform reconstruction only

on the results, we focus our attention on threshold schemes in which s is reconstructed directly. A well-known such construction, due to Shamir [200], distributes points lying on a polynomial (of degree $t-1$) as shares. We refer to this construction (summarized in Appendix A.1) as *classic Shamir secret sharing*. Combining t of these shares using polynomial interpolation would recover the secret; combining any smaller number of shares does not leak any information about the secret. The construction is information theoretically secure; that is, a Shamir threshold scheme can withstand adversaries with unlimited computational power.

4.3 Variability of Threshold Schemes

Threshold schemes allow for a high degree of variability in user goals and potential adversaries, where even slight variations significantly influence the security of the scheme overall. We describe two practical examples, where both cases utilize an identical underlying threshold scheme protocol, however, the threat model, context, goals, and thus ceremony vary dramatically between the two examples. For both examples, Alice, Bob, and Carol are journalists at the same organization.

Case One. Alice received highly sensitive files from a source. She fears external parties will act against her to prevent the distribution of the files, and wants to ensure that even if an adversary succeeds at targeting her, the information can still be accessed by either herself or trusted colleagues. She enlists the help of Bob and Carol in her efforts to preserve the availability of the files.

Alice acquires a laptop to encrypt and store the secret information. She inputs a key k into a tool that implements classic Shamir secret sharing using k as the secret ($s = k$) and inputs her desired parameters $t = 2$ and $n = 3$. The tool outputs the corresponding shares and Alice messages Bob and Carol over an established communication channel. Alice sends one share to Bob and one share to Carol. Bob and Carol confirm they received the share and each store their respective shares on a USB, which is then stored in a chosen safe place. Alice stores the laptop containing the encrypted secret information in a safety deposit box at a bank.

Alice leaves the news organization and Bob, who has lost his share, is assigned the story. Fortunately, Alice left the organization on good terms, so Bob contacts Alice and Carol, requesting their shares. Bob retrieves the laptop and decrypts the ciphertext using the key recovered from Alice and Carol's shares.

Case Two. Alice has received a decryption key that corresponds to a publicly released

ciphertext [15, 16]. She fears external parties will attempt to distribute the key in a way that could endanger individuals. Alice wants to ensure the information remains confidential to those she has not authorized (herself or her trusted colleagues Bob and Carol) to distribute it.

Alice meets Bob and Carol at a previously agreed upon location. Using an airgapped[1] laptop, Alice inputs the key from her source into a tool that implements classic Shamir secret sharing, choosing $t = 2$, and $n = 3$. After the tool has output the corresponding shares, Alice, Bob, and Carol each save one share to their respective USB. Finally, Alice deletes all information off of the airgapped laptop. Everyone keeps their respective USB devices on their person at all times.

Alice's USB is taken from her while crossing a border. Fortunately the USB is insufficient to learn the secret data, however, Alice can meet Bob and Carol in person to request their shares in order to recover the key.

Observations. In both cases, Alice made a number of choices, including selection of participants, selection of communication methods, and selection of storage mechanisms. The choices Alice made affect the security and privacy properties of each case. For instance, only Case Two requires the physical presence of each participant. Such a requirement may limit the availability of the information if an adversary had the power to prevent participants from meeting up; for example, if the participants are initially separated by a geographical border. Furthermore, storing the encrypted data on the laptop creates a single point of failure for an adversary targeting the availability of the ciphertext.

The above examples demonstrate how the range of choices and prioritization impact the threshold scheme ceremony. For instance, Case One preserves availability and prevents the information from being released preemptively, but still requires confidentiality, otherwise multiple copies of the information could simply be stored. More significantly, these examples highlight the need to consider the ceremony of the threshold scheme when performing a security analysis, factoring in both the threat model that users operate within, along with how users perform the actions required of the threshold scheme in context of their use case, such as whether users operate online or entirely offline. This crucial observation motivates our next sections, where we formalize several ceremonies of threshold schemes as both protocol *and* explicit user actions and decisions in the effort to more accurately assess the security of the threshold scheme under consideration.

[1] As a security mechanism, an airgapped laptop is protected against connecting to the Internet.

4.4 Related Work

Ceremony Analyses of Security Tools. Our analysis follows in the style of prior work analyzing cryptographic tools used in practice [59, 78], verifying security properties, and documenting decision paths taken by users when participating in cryptographic protocols. Our security analysis specifically includes the ceremony performed by end users [64], encompassing everything out-of-band to a cryptographic protocol but required of users and thus subject to security consequences [56, 102].

While prior ceremony analyses have been performed for a number of cryptographic protocols, the research literature currently lacks a similar analysis assessing the security of ceremonies for threshold schemes specifically. Previous work on ceremony analysis includes specifying how to model users' devices [137] and formal analysis of Public-Key Infrastructures (PKIs) [138]. While frameworks have been proposed to assess the security of existing ceremonies such as that of Carlos et al. [33], our framework specifically defines possible threshold scheme ceremonies and facilitates their analysis.

In 2013, Carlos et al. [34] focused on threat modeling within ceremonies, highlighting that threat models for ceremonies must be adaptive. Threat models must evolve to match varying user goals and contexts even when these ceremonies utilize the same underlying protocol. Radke et al. [181] highlight adaptive threat models as a weaknesses of ceremony analysis, as the context of the ceremony must be as well defined in order to accurately model potential adversaries and threats against the goals of end users to claim the ceremony as secure. As ceremony analysis of threshold schemes is highly dependent on users' threat models, and as the threat models can differ depending on the user, context, or use case, instead of providing a narrow ceremony analysis for a specific threat model, we present a generalized *framework* for performing a ceremony analyses of threshold schemes in both theory and practice, across several common real-world threat models.

Applications of Threshold Schemes. Sunder is an existing applied secret sharing tool created by Freedom of the Press Foundation [77] to support journalists protecting long-term secrets such as the Snowden archives. Another tool building on secret sharing is Callisto, which provides a safety-in-numbers approach to exposing names of sexual abusers [182]. While these use cases give insight into the setting and application of threshold schemes used in practice, many other use cases of threshold schemes exist [12, 13, 218].

Shatter [13] is a framework for desktop and mobile platforms that performs key sharing across a user's devices. Shatter uses secret sharing to leverage users' increasing numbers of devices by requiring a threshold number of devices to provide consensus for actions such

as performing a login. Although Shatter uses secret sharing it is actually an example of a threshold signature scheme. Nonetheless, it still shares a number of properties with secret sharing schemes that make our analysis applicable.

Shatter Secrets [12] is an advance on Shatter that provides protection to users' data when crossing borders. With Shatter Secrets, a user could encrypt their primary device and then distribute shares to their friends at their destination with the encryption key serving as the secret s. Once over the border, the user with the encrypted device visits t of their friends, physically NFC-taps their devices to retrieve the shares, reconstructs the secret, and decrypts their device.

Pico [218] stores shares on hardware tokens instead of utilizing users' existing devices. One explicit use case of Pico is as a replacement for password managers as it uses public key cryptography challenge-response instead of typical passwords. Pico exists as a mobile application and is intended to block a thief who has stolen fewer than t tokens from violating confidentiality, while preserving availability as long as t tokens remain.

Secret Sharing Variants. Verifiable Secret Sharing (VSS) is a variant on threshold schemes in which any participant can verify the integrity of their share using a public commitment. Well-known VSS schemes include Feldman's [68] and Pedersen's [171] schemes.

Proactive secret sharing, introduced by Ostrovsky and Young [166] and used in a secret-sharing scheme by Herzberg et al. [92], protects against a *mobile adversary*. A mobile adversary can control a subset of players over time, but the members belonging to this subset can change between epochs. To defend against such an adversary, proactive secret sharing relies on proactively updating shares to enable a form of forward security.

4.5 A Framework for Ceremony Analysis

In this section we present the components we need for performing a ceremony analysis. We define threshold scheme adversaries, distinguish two modes of operation, identify security goals, and provide additional terminology that we use throughout our analysis. We conclude this section with an outline of how to use our framework to produce a complete ceremony for a specific secret sharing protocol used in a specific setting and purpose.

4.5.1 Formalizing Threshold Scheme Adversaries

First, we formalize a range of possible adversaries against threshold schemes used in practice, and describe the possible capabilities and powers these adversaries can hold. We

outline several conventional adversarial models and identify variations within each model.

Adversary Power. We define three levels of power an adversary may possess. Although we utilize the terms 'high', 'middle', and 'low', these terms are simply points of reference for comprehension and are not intended as prescriptive classifications.

A *high-powered adversary* has the power and resources of a government actor. High-powered adversaries can access state-of-the-art computing resources and have significant quantities of time and money at their disposal. Such an adversary has the power to take legal action, bounded only by the political environment of that jurisdiction. For example, the NSA is known to masquerade as well-known sites, installing malware capable of exfiltrating data from a victim's device [80], governments are known to use informants to infiltrate activist groups (Martin Luther King Jr.'s friend and photographer was an FBI informant [142]), and some countries have proposed laws allowing legal orders requiring technology companies to work on behalf of the government to provide access to encrypted devices [55].

A *low-powered adversary* has similar computational, temporal, and monetary resources as the participants of the threshold scheme. A *middle-powered adversary* exists somewhere between the powers of a government actor and the powers of the participants. Such an adversary has the same legal powers as a low-powered adversary, but may have the same money and time available to them as a government actor.

Adversary Capabilities. We limit our analysis to the capabilities of the below-mentioned adversarial models. The adversaries may be participants in the ceremony or outsiders. A previously trusted participant may become an adversary at a later time in the protocol. That is, we do not assume participant roles are static.

An *honest-but-curious* (HBC) adversary will not deviate from the ceremony, but will try to learn as much information as they can within the bounds of the ceremony. An HBC adversary will view any information that is exposed to them, and may collude with other participants in an effort to learn additional information.

A *malicious* adversary is not bound to any expectation of behaviour, and she can participate both honestly and dishonestly in the ceremony at will. A malicious adversary can impersonate other actors, elect to not participate in the ceremony, or participate disruptively by, for example, providing false shares and attempting to deceive other parties into providing the adversary with their shares.

Adversaries who compromise operating systems or hardware infrastructure are also a real threat to users defending against a high-powered adversary. However, this class of threats are out of scope for our analysis as details of physical infrastructure vary widely

between implementations. In practice, secret-sharing implementations should employ well-known device protection techniques such as single-use strategies[2] and using secure operating systems such as Qubes [197].

Adversary Goals. Here we identify goals for an active adversary of threshold schemes.

Learning secret information. An adversary motivated to learn the sensitive information $\mathcal{F}$ can work to gain knowledge of the secret or the shares, subsequently allowing the adversary to recover $\mathcal{F}$.

Modifying secret information. An adversary may wish to modify $\mathcal{F}$ without detection, resulting in participants recovering information that is different than the original input into the threshold scheme.

Preventing secret recovery. Adversaries may also seek to prevent others from accessing or disseminating $\mathcal{F}$. For example, an adversary seeking to hide information—such as a government seeking to prevent public distribution of evidence of war crimes—can work to disrupt communication, destroy shares, or even to destroy the sensitive data $\mathcal{F}$.

Causing harm to participants. In some countries, working with material that is prohibited can be a crime, putting all parties at risk [19]. An adversary may be motivated to harm the participants of the threshold scheme, and can seek to perform actions such as attributing ownership of $\mathcal{F}$ to those participants.

4.5.2 Modes of Operation

We define two manners of use, termed 'modes of operation', to manage the sensitive information $\mathcal{F}$. The Base Mode is defined as a ceremony for classic Shamir secret sharing [200]. The Extended Mode is an extension to Base Mode, and is documented to be used in practice in high-risk settings [12, 77].

Base. In the first mode of operation, the confidential information, $\mathcal{F}$, is small in size, such that each of the shares distributed to the participants can reasonably be about the size of $\mathcal{F}$. The secret s can then be the information itself, $s = \mathcal{F}$.

Extended. The second mode of operation, addresses when the sensitive information is too large to be used directly as the secret s. The **Ext** mode of operation is modeled after a common real-world use case described in the documentation of the secret sharing tool Sunder [77]. In this case, the confidential data $\mathcal{F}$ is first encrypted and the encryption key

[2]One example of a single-use strategy is using "burner" phones. A burner phone is one that is newly purchased and used for a short period, after which it is discarded. [196]

is then used as input as s into **Base**. After the secret s is reconstructed, additional steps must be taken to retrieve the data $\mathcal{F}$ using s as a key.

4.5.3 Identifying Security Goals

We next define several security goals that are commonly cited for implementations of threshold schemes used in practice [13, 77]. Later, we assess these security goals to determine the extent to which these goals are achieved in several common real-world settings and across a range of implementations and commonly used schemes, as well as in our identified improvements detailed in Section 4.9. The below identified goals are specific to the context and use case of threshold schemes in general. However, the context of the threshold scheme under consideration can impact the security goals for the scheme. We use this set of security goals for our analysis, but other analysts using our framework should identify the security goals appropriate for their specific scheme. Such goals are not limited to the ones listed here and may include some of these goals, and others as well.

1. **t-Separation of Privilege:** We define *t-Separation of Privilege* as a specific case of the well-known *Separation of Privilege* security principle first introduced by Saltzer and Schroeder [193]. Threshold schemes require t participants' shares to perform a recovery of $\mathcal{F}$, where t is the chosen threshold.

2. **Availability:** The secret information $\mathcal{F}$ is accessible to honest participants so long as at least t valid shares remain accessible. For Extended Mode the availability of the encrypted version of $\mathcal{F}$ will be enforced by the choice of safe storage mechanism (see Section 4.5.5).

3. **Information Theoretic Security:** Even given unlimited computational power, adversaries inside or outside the ceremony cannot access $\mathcal{F}$ while possessing fewer than t shares.

4. **Confidentiality:** Adversaries outside of the protocol cannot gain knowledge of $\mathcal{F}$. Note that in a real-world setting this goal requires revocation of participants to achieve confidentiality across epochs where participants move from a trusted to an untrusted state.

5. **Integrity/Corruption Detection:** Corruption of an individual share or the sensitive information is detected by honest participants before completing the Reconstruction stage.

4.5.4 Threshold Ceremony Analysis Outline

We next describe our framework to structure our assessment of the security of threshold schemes in practice, including both the protocol and ceremony of the threshold scheme under consideration.

Identify stages of the ceremony of the threshold scheme. Security ceremonies can be broken down into components called *stages*. Fully specifying the complete ceremony and its component stages is the first step towards evaluating the security of the threshold scheme under consideration. We provide two formalizations (Sections 4.6 and 4.7) as a skeletal frame of reference for future analyses of threshold schemes derived from Shamir secret sharing.

Define the threat model. First, define possible *adversaries* of the threshold scheme or of the users participating in the scheme, including the adversaries' goals. In Section 4.5.1, we demonstrate a range of possible adversaries against threshold schemes. Second, determine the desired *security goals*. We present several possible security goals of threshold schemes in Section 4.5.3. At times, certain security goals may prove to be in conflict. For example, a system operating in Extended Mode that prioritizes availability over confidentiality might distribute an encrypted ciphertext publicly in order to decrease the possibility of destruction.

Define the mode of operation. Threshold schemes can potentially allow for many modes of operation. For example, classic Shamir secret sharing can support both the Base Mode and Extended Mode of operation. To evaluate the security of a threshold scheme, a single mode of operation must first be specified. If a scheme supports more than one mode of operation, the security evaluation should be performed once for each. Note that transitioning between modes of operation for the same, or updated, secret is not supported by such evaluations as it introduces new potential attack vectors (including issues related to using shares at most once; see Section 4.5.5).

Evaluate security goals against adversaries. Using knowledge of adversary goals and capabilities along with the ceremony formalization, the security goals for the system can be evaluated in the context of the given mode of operation and threat model. For each stage in the threshold scheme, and for each step within a stage, evaluate if the adversary's capabilities can defeat the system goal. If the adversary can defeat the system goal, this goal is considered unmet. See Figure 4.1 for an overview of our framework.

1. Ceremony identification and formalization (stages)
2. Threat Model (selection of adversaries and security goals)
3. Mode of operation (identification of use cases)
4. Evaluation of Security (assessment of security goals relative to threat model)

Figure 4.1: Framework for security analysis for threshold schemes derived from Shamir secret sharing

4.5.5 Assumptions and Limitations

We maintain several assumptions for the purpose of providing a structured analysis and designing our framework. However, we acknowledge these assumptions may not always hold in real-world settings.

Secure Communication and Storage. We emphasize the existence and availability of a secure communication channel as well as a mechanism for safe storage. Communicating and storing data securely are both critical to the security of a practical threshold scheme.

Safe storage is a storage mechanism such that data is guaranteed to be recoverable in the future. Such mechanisms must avoid single points of failure such as due to server crashes; preventing such failures requires storing copies of the data on multiple servers, for example. We assume a safe storage mechanism provides the properties of availability to participants. We also assume that if participants require authentication before accessing the stored data, the safe storage mechanism can provide this authentication mechanism.

Using Shares At Most Once. We maintain that secrets, and consequently shares, should be single-use as otherwise new security risks are introduced. For example, a multi-use setting requires the assumption that the participant performing the recovery securely deletes both the secret and the collected shares from their local device *after* completing the recovery. If the recovering participant breaks this trust and stores shares or the secret information locally after the first recovery, the participant can bypass the step of gathering $t-1$ shares from other participants in future recoveries. Furthermore, a device containing t shares is a single point of failure—an appealing target for an adversary trying to learn the secret.

Honesty of the Dealer. We assume that the dealer is honest both in the case of D and D_P, as classic Shamir secret sharing is trivially broken when the dealer is dishonest.

In the Extended Mode the dealer is also responsible for determining sufficient protection for the encrypted $\mathcal{F}$ in terms of the secure storage selected.

Erasure Assumption. For the purposes of our analysis, we work within the erasure assumption [31], which assumes that participants are able to securely erase data when required. We recognize that if the erasure assumption does not hold, then many of the security properties we define are broken, as an adversary could perform analysis post-hoc on stolen machines and recover sensitive material.

Non-Collusion. Honest participants will not collude with external parties. For example, honest participants will not attempt a recovery initiated by an unauthorized person.

4.6 Base Mode Stages

We now more formally identify the possible choices and actions for users participating in a threshold scheme, and introduce a formalization for a general ceremony of threshold schemes based on Shamir secret sharing. Starting with the Base Mode of operation, we present three stages consisting of *share generation*, *share distribution*, and *reconstruction*. The ceremony framework for the Base Mode of operation is outlined in Figure 4.2.

Classification of Steps. We annotate each step in a stage to classify how participants are involved. We annotate steps as *Device* for expected implementation actions, as *Choice* for user decisions, and as *Action* for expected user actions.

4.6.1 Share Generation

The generation stage allows minimal variation and choice from the user. In this stage, we assume a secret s has previously been selected. A dealer D possesses s and selects the values for t and n. The dealer may or may not be a participant in the scheme. Regardless, the dealer provides t, n, and s to a tool that follows the steps for Share Generation defined in Appendix A.1, resulting in the generation of n shares. After these shares have been generated, the device should securely delete all r_i's (which it created) while the dealer deletes all copies of s.

The choices and actions required of the dealer at this stage consist of selecting appropriate values for t and n.

Choice: Determine Parameters. When generating shares, the dealer chooses the appropriate threshold and number of participants. As these choices are highly context

Share Generation

1. Choice: The dealer chooses values for n and t.
2. Device: Let the secret space be $\mathcal{S} = GF(q)^{\ell}$, where q is a prime or a prime power, $q \geq n+1$, and $\ell \geq 1$. Let $s \in \mathcal{S}$ be the secret.
3. Device: Selects $t-1$ values independently and uniformly at random from $\mathcal{S}$ as $r_1, \ldots, r_{t-1}$ and sets $f : GF(q) \to \mathcal{S}$ as $f(x) = r_{t-1}\, x^{t-1} + r_{t-2}\, x^{t-2} + \cdots + r_1\, x + s$.
4. Device: Generates shares $s_i = (a_i, f(a_i))$ for $1 \leq i \leq n$, where the a_i are arbitrary distinct non-zero elements of $GF(q)$.
5. Device: Delete r_i's.
6. Action: Delete all copies of s.

Share Distribution

1. Choice: Select n participants (possibly including the dealer).
2. Choice: Select a secure communication channel (in person, Signal, etc.).
3. Action: The dealer distributes $s_i = (a_i, f(a_i))$ to participant P_i for $1 \leq i \leq n$.
4. Action: Delete each s_i from the dealer's device. Exception is if the dealer is a participant and keeps one share.
5. Choice: Each participant selects an appropriate storage mechanism for their share.
6. Action: Each participant stores their share in the selected storage mechanism.

Reconstruction

1. Choice: Select a communication channel to bring t or more shares together.
2. Action: P_r and the contacted participants authenticate one another.
3. Choice: Contacted participants elect whether to proceed and participate in a reconstruction.
4. Action: If proceeding, a contacted participant sends their share to P_r.
5. Device: Combine the t or more shares using polynomial interpolation to recover the secret $s = f(0)$.

Figure 4.2: Ceremony Framework for Base Mode of Operation

dependent, the dealer is trusted to make these choices taking into account their respective threat model.

An adversary in this setting can leverage poor or uninformed choices of n and t to gain unauthorized access to or prevent participants from accessing s. For example, an adversary hoping to prevent a group of journalists from accessing the sensitive information need only destroy $x > n - t$ shares to prevent journalists from accessing the sensitive information in the future.

Choosing t and n requires identifying the trade-off in prioritization for availability and t-separation of privilege or risk of collusion. A larger value for t (for fixed n) increases the number of participants that can collude without learning the secret, while lower t increases the number of participants that can be unavailable while still keeping the secret in a recoverable state. The value of n, on the other hand, is likely to be determined by context—specifically by how many trusted participants are available, as opposed to n being easily chosen.

Action: Perform Secure Deletion. Secure deletion is always required when performing share generation. Verification that secret material has been securely deleted[3] is difficult, thus for our analysis we work under the assumption of the erasure mode defined in Section 4.5.5. Achieving the desired security properties of the threshold ceremony requires the dealer to delete s and all r_i's after generating shares. This fact demonstrates the higher level of trust required in the dealer beyond that of the other participants.

If the dealer fails to delete s and the r_i's off their machine it becomes an easy and highly profitable target for an adversary. This single point of failure allows the adversary to bypass reconstructing t shares and instead target the dealer's machine. Thus, the presence of s on the machine of the dealer presents a formidable risk and underscores the necessity of secure deletion.

4.6.2 Share Distribution

Share distribution determines who receives shares and how the shares are transmitted to participants. The responsibility of the dealer includes selection of participants, selection of a secure communication channel, and transmission of shares over this channel.

After receiving their share, a participant is responsible for selecting a safe storage mechanism for the share until required for the Recovery stage. After all shares are distributed,

[3]For further details on existing secure deletion solutions see the analysis from Reardon [183].

	Confidentiality	Integrity	Info. Theor. Sec.
In-person	●	○	●
Signal	●	●	○
TLS 1.3	●	●	○
PGP	◐	●	○
SMS	○	○	○

Table 4.2: Network Model: Properties of Communication Channels
●=achieved; ◐=potential loss; ○=not achieved

the dealer securely deletes all shares from their device, with the exception of their own share, if applicable.

Choice: Select Secure Channel. Shamir secret sharing assumes the existence of a secure communication channel. However, the dealer holds responsibility to assess and choose an appropriate channel where all aforementioned security properties hold. Unsurprisingly, users often struggle to make safe choices when using security-critical tools in similar contexts [232]. Communication channels that could be used in practice which are not information-theoretically secure include TLS [189], Signal [172],and PGP [29], while in-person communication achieves information theoretic security. Notably, each of these transmission methods achieve divergent security properties when used in a secret sharing ceremony. TLS and Signal support confidentiality and integrity assuming that participants authenticate one another before sending any messages. In-person communication achieves confidentiality but does not support integrity, due to the lack of a defined integrity mechanism. While PGP encrypts data in transit, thereby achieving confidentiality at the moment data is transmitted, PGP is not forward-secure. Consequently, PGP does not strictly preserve confidentiality of future transmitted data in the case that a user's private key is compromised. Cellular networks' Short Message Service (SMS), while commonly used for security protocols such as two-factor authentication [103], does not achieve any of our desired properties. We provide a summarized analysis of the security properties of various channels in Table 4.2.

Choice: Select Participants. In a (t, n)-threshold scheme, the dealer is responsible for selecting which participants are entrusted with shares. The dealer in some cases is also free to decide if they will become a participant dealer D_p and retain a share for themselves. Choosing appropriate participants is heavily context dependent and influenced by the users' threat model.

Action: Perform Secure Deletion. As with Share Generation, there is a need for secure deletion. After distributing the shares the dealer must delete all shares that are not their own from their device. Failure to do so provides an adversary with a single target to gain the secret s just as leaving the secret itself would.

4.6.3 Reconstruction

This stage occurs when a valid participant chooses to initiate a recovery. The participant P_r performing the recovery contacts and authenticates other participants, who authenticate P_r as a valid participant. These participants then decide for themselves whether reconstruction is appropriate and whether to participate at that time. If so, they transmit their share over a secure channel. Once P_r possesses t or more shares, P_r can perform reconstruction using a tool for polynomial interpolation to extract the secret s.

Choice: Select Secure Channel. Performing a recovery of s assumes a secure communication channel to transmit shares from other participants to P_r.

Action: Perform Authentication. During recovery it is left to the participants to authenticate each other even assuming a secure channel. For instance, the participants in the scheme need to know whether or not it is permissible for the initiating participant to perform a recovery. In one example, from Section 4.3, Alice left the organization and Bob determined it was okay to include Alice in the recovery and contacted her. However, in another setting we can imagine Alice left the organization and initiated a recovery by requesting a share from Carol. Without a revocation mechanism, there is nothing preventing Alice from recovering the secret if Carol provides her with a share. Therefore, even after losing authorization, Alice can learn the secret and break confidentiality. Thus, in this latter setting, the ceremony as stated is insecure. Such an insecurity is an example of how a ceremony secure in one case may be insecure in another, and so it is important to specify the ceremony as part of the security analysis, as opposed to just analyzing the underlying protocol. Additionally, we will address this particular insecurity in Section 4.9.

4.7 Extended Mode Stages

The Extended Mode is one possible extension of Shamir secret sharing, and is a practical use case for users seeking to protect sensitive information that is large in size. For example, Sunder requires operating in the Extended Mode when the secret is larger than 1 MB [77].

Secret Preparation

1. Device: Generate a secret key to be used as s.
2. Device: Encrypt $\mathcal{F}$ using s and an appropriate authenticated encryption algorithm.
3. Choice: Select a safe storage mechanism for the ciphertext.
4. Action: Safely store the ciphertext.

Extended Reconstruction

1. Action: Acquire ciphertext from selected safe storage.
2. Device: Use recovered s to decrypt the ciphertext.

Figure 4.3: Ceremony Framework Additions for Extended Mode of Operation

We now formalize the choices and actions users must make in the Extended Mode. We introduce the stages *Secret Preparation* which is performed before the Base Mode Share Generation stage, and *Extended Reconstruction* which is performed after the Base Mode Reconstruction stage; see Figure 4.3 for an outline.

4.7.1 Secret Preparation

Secret Preparation begins with an existing plaintext and a secret key. Using this key, the plaintext is encrypted via a symmetric encryption algorithm, and the output of the ciphertext is stored using safe storage (see Section 4.5.5) for later use in the Reconstruction stage. This secret key is subsequently used as the input s into the Share Generation phase.

Action: Generate Secret Key. The sensitive information $\mathcal{F}$ should be encrypted using the chosen authenticated encryption algorithm and a secret key generated by the dealer. Note that authenticated encryption does *not* provide end-to-end integrity against an attacker that is able to acquire s. In that event, an adversary could modify the stored ciphertext without detection. We present a way to block this attack in Section 4.9.

Choice: Select Safe Storage. Even if s has been successfully reconstructed, the user must have chosen a reliable storage mechanism (see Section 4.5.5 for requirements of safe storage) to recover the ciphertext of $\mathcal{F}$ after performing the Reconstruction stage.

4.7.2 Secret Recovery

After s is recovered in the Reconstruction stage, the participant initiating the recovery can retrieve the ciphertext of $\mathcal{F}$ from the chosen storage location, use the secret s as a key for the symmetric encryption algorithm for decryption, and produce the original sensitive information $\mathcal{F}$.

For this stage, the probability of successfully recovering $\mathcal{F}$ is dependent on choices made by the user in the Secret Preparation stage; for example, if the user did not choose an adequate storage mechanism, the user may fail to recover $\mathcal{F}$ in the Secret Recovery stage.

4.8 Application of Ceremony Framework Analysis

We now apply our ceremony framework to aid the security analysis of threshold schemes, as specified in Section 4.5. We present three case studies to highlight how seemingly straightforward implementations can achieve or miss assumed security goals.

4.8.1 Defined Threat Model

We maintain a specific threat model for our security analysis of each case study.

Adversaries. We assume a high-powered adversary (defined in Section 4.5.1) which has access to exceptional computational power, time, and money, along with significant legal resources. We do not assume fixed roles for participants; a once-trusted participant can become an adversary at a later time.

Security Goals. We evaluate each case study against the sample security goals defined in Section 4.5.3. Specifically, we evaluate the security goals of t-separation of privilege, availability, information theoretic security, confidentiality, and integrity against the above-defined adversary.

4.8.2 Case Study One: Classic Shamir Threshold Scheme

A summary of the analysis detailed below for classic Shamir secret sharing (and indeed all of the ceremonies we analyze) can be found in Table 4.3. Classic Shamir secret sharing is

	Classic				Sunder Ceremony				Shatter [12]		Ours			
	Base		*Ext*		*Base*		*Ext*		*Ext*		*Base*		*Ext*	
	HBC	MAL	HBC	MAL	HBC	MAL	HBC	MAL	HBC	MAL	HBC	MAL	HBC	MAL
t-Sep. Priv.	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Availability	●	●	○	○	●	●	◐	◐	○	○	●	●	●	●
IT Sec.	◐	◐	○	○	○	○	○	○	○	○	○	○	○	○
Conf.	◐	◐	◐	◐	◐	◐	◐	◐	●	●	●	●	●	●
Integrity	○	○	○	○	●	●	◐	◐	◐	◐	●	●	●	●

Table 4.3: Ceremony Analysis Summary, note IT-Sec=Information Theoretic Security
●=achieved; ◐=ceremony dependent; ○=not achieved

not a complete protocol and by extension not a complete ceremony. Unsurprisingly, classic Shamir secret sharing in isolation cannot achieve all desired security properties.

t-Separation of Privilege is always achieved by classic Shamir secret sharing. Even within the Extended Mode, an adversary with access to the ciphertext still requires at minimum t shares to decrypt the ciphertext. In contrast to the above, the loss of *Availability* in Extended Mode demonstrates how security properties can be lost when moving from Base Mode to a seemingly innocuous extension of Shamir secret sharing. In the Base Mode, availability is preserved as long as $t < n$. In the Extended Mode, the loss of the ciphertext renders the secret unavailable, regardless of the number of shares that remain available. As the protocol does not define a safe storage mechanism to protect against loss of the ciphertext, this becomes a single point of failure.

Information theoretic security is achieved in theory by the mathematics of classic Shamir secret sharing. When evaluating the scheme in practice, we must consider the channel used to transmit shares to participants. We grant a half-circle in the table for information theoretic security in Base Mode as the protocol can remain entirely offline if desired, requiring shares to be transmitted in person or via a trusted physical channel. However, when operating in online mode, shares are transmitted over an online channel. As online communication channels rely on encryption protocols that are not information theoretically secure, classic Shamir secret sharing loses information theoretic security when used with an online channel. Furthermore, as Extended Mode requires a symmetric encryption algorithm, working within the Extended Mode similarly is not information theoretically secure.

Confidentiality is not achieved in either the Base or Extended Mode of operation due to the lack of a revocation mechanism. Once shares have been distributed, classic Shamir secret sharing does not consider the case where a once-trusted participant moves to an untrusted state, such as by voluntarily or involuntarily leaving an organization. For example, nothing prevents a participant who was fired, but possesses a valid share, from participating in a future recovery protocol by colluding with other participants who similarly may or may not be currently within a trusted state in the organization. We therefore only grant half-circles in the table for all of these cases, as they only achieve confidentiality as long as there is no need for revocation.

Integrity of shares is not a goal that classic Shamir secret sharing guarantees. In some settings, for example, $t = 2$ but four shares are available during the reconstruction phase, the correct secret can be determined if a limited number (in this case, one) of shares is corrupted or maliciously changed. However, using these techniques for integrity requires raising the required number of shares during reconstruction, as well as assumptions about

the number of corrupted shares. Ideally, a separate integrity check for the shares would enable the detection of corrupted shares directly. Once detected and identified, corrupted shares would be excluded from the recovery and, if necessary, additional validated shares could be included; we provide this functionality in our extensions in Section 4.9. Note that in the Extended Mode the ciphertext carries its own integrity check, but that check is not entirely sufficient, as discussed in Section 4.7.1.

4.8.3 Case Study Two: Sunder

Freedom of the Press Foundation's tool Sunder [76, 77] is a desktop application for journalists to generate a configurable number of key shares for encrypted documents. While this tool supports both Base and Extended Modes of operation and can accommodate a wide range of threat models and ceremonies, we bound our analysis to the online setting with a high-risk threat model and high adversary capabilities. We define an explicit ceremony for Sunder in Appendix A.2.

Sunder is a straightforward implementation of classic Shamir secret sharing, and many of the security properties for classic Shamir secret sharing apply to Sunder. In the secret-sharing implementation used by Sunder [217], every character in the ℓ-character secret provided to Sunder becomes an input into a Shamir protocol acting over the Galois Field $GF(256)$. (Equivalently, the secret as a whole is treated as an element of the vector space $GF(256)^{\ell}$.) Each of the n shares will then be ℓ bytes long. One deviation worth highlighting is Sunder's support for share integrity. The underlying cryptographic library [217] that Sunder utilizes provides share integrity by generating a public-private ephemeral key pair to sign shares during the Generation stage. Share signatures are validated during the Recovery stage to ensure both the validity of shares and also that all shares are signed by the same public key. However, Sunder does not include document encryption within the tool and thus does not support integrity validation for the encrypted documents; consequently, Sunder only partially attains integrity for Extended Mode.

Sunder can achieve confidentiality if no more than t participants holding valid shares leave the organization, however, Sunder is limited to confidentiality without revocation and thus only has a half-circle in the table.

Finally, while Sunder provides availability in Base Mode, it leaves to the user the decision of how to store the encrypted file in Extended Mode. If the file is not safely stored, availability could be compromised.

4.8.4 Case Study Three: Shatter Secrets

Shatter Secrets [12] is an open-source protocol that uses Shamir secret sharing of a key that encrypts a user's device. Shatter secrets can be used to distribute shares (each encrypted with a key held by the user) to other devices and friends of the device owner such that a threshold number is required to decrypt the device. A device owner crossing an international border can encrypt their device using Shatter Secrets such that they are unable to decrypt their device without the physical presence of a threshold number of participants holding shares. We define an explicit ceremony for Shatter Secrets in Appendix A.3.

The shares in Shatter Secrets are themselves encrypted by a key held by the primary data owner (on a secondary device). Violating confidentiality thus requires compromising the encrypted primary device, the secondary device storing the share decryption key, and a sufficient number of the friends' devices storing the encrypted shares. Specifically, unlike in Sunder, shareholders alone cannot recover the secret. Thus confidentiality of the device's data is preserved. Availability of the device's data, however, can be easily compromised as authorities can seize the encrypted device (assuming the device owner does not have a backup). Shatter Secrets' design does not adequately provide the property of availability as it prioritizes confidentiality such that the device can be a single point of attack. The authenticated encryption of the shares themselves provide integrity for the shares, but the integrity of the encrypted device depends on whether that encryption mechanism can detect modifications to the ciphertext.

4.9 Lightweight Integratable Improvements

As can be seen from our example case studies, implementations based on classic Shamir secret sharing have gaps limiting the security properties, such as confidentiality and integrity in Base Mode and availability in Extended Mode. To address these gaps, we introduce a lightweight set of improvements which are fully compatible with classic Shamir secret sharing. These improvements are extensions to Shamir secret sharing and can be applied to implementations of Shamir secret sharing. Using our framework for ceremony security analysis, we assess the security properties provided by these improvements within the Base and Extended Modes.

4.9.1 Overview

We define a lightweight Proactive Verifiable Secret Sharing (Proactive VSS) scheme and specify three new stages: Share Update, Share Validate, and Generate Commitment. These stages provide participants with the capability to update shares and revoke access to individuals who are no longer trusted. Users can verify the integrity of shares and verify the integrity of $\mathcal{F}$.

We use the protocol and model of Herzberg et al. [92], in which adversarially controlled players during an update stage count against both adjacent epochs. Alternatively, one could use the protocol and somewhat stronger adversarial model of Nikov and Nikova [157], at the cost of requiring the dealer to select t knowing that $t-1$ corrupted players could compromise the secret, but t are needed to reconstruct it.

Assumptions. Our commitment scheme assumes a sufficiently random s, such as a key. If s is not sufficiently random, several of our improvements can still be used; see Section 4.9.4.

4.9.2 Base Protocol Description

Our modifications can be used in conjunction with an existing secret sharing implementation as demonstrated by the modifications to the stages indicated in Figure 4.4.

Modified Share Generation. In addition to the original steps, the dealer $\mathcal{D}$ generates a commitment $\vec{C}$ to the polynomial. The j^{th} index of $\vec{C}$ is equal to g^{r_j}, where r_j is the randomly selected coefficient, while the zeroth element in $\vec{C}$ is equal to g^s. (Here, g is the generator of a group in which the Decisional Diffie-Hellman problem is hard.) The dealer publishes $\vec{C}$ to a trusted location that every participant can access. Examples of such a location include a commitment verification party, a public blockchain, or some other location all participants can reliably view in the same state. Note that the choice of trusted location is influenced by the powers of an adversary—a trusted Twitter account could be effective against a low-powered adversary, but not a high-powered adversary.

Share Validation. Once $\vec{C}$ has been published, any participant P_i can validate the integrity of her share $s_i = (a_i, f(a_i))$, where $f(a_i)$ is the value generated for participant i using the assigned value a_i and the dealer-selected function f. Validation requires first fetching the public commitment and computing $\psi = \prod_{j=0}^{t-1} \phi_j^{a_i{}^j}$ where $\phi_0 = g^s$, and $\phi_j = g^{r_j}$ for $1 \leq j \leq t-1$. Next, each participant validates that $g^{f(a_i)}$ is equal to ψ, where $f(a_i)$ is taken from their respective share.

Share Updates. Share updates use commitments to zero to preserve the original secret value upon secret reconstruction. To perform an update on m shares, where $t \leq m \leq n$, the participant performing the update assumes the temporary role of the updater $\mathcal{U}$, where $\mathcal{U}$ can be any valid participant. If $m < n$, then there will be shares that do not receive an update and therefore are effectively revoked.

The updater $\mathcal{U}$ first generates m share updates and one commitment update. The set of share updates is generated by running the Share Generation stage with $s = 0$. (Let the polynomial used in this stage be $h(x)$.) This step ensures that shares from a prior epoch cannot be used in conjunction with updated shares in the next epoch to reconstruct s. Thus, shares can be proactively "rotated" forward while protecting the original secret.

After Generation, $\mathcal{U}$ distributes the m share updates to the selected participants. Additionally, $\mathcal{U}$ must apply the commitment update to the original commitment by performing pointwise multiplication between the original commitment to $f(x)$ and a new commitment to the new polynomial $h(x)$. Consequently, $\mathcal{U}$ must be able to safely access the commitment such that the commitment update can be securely applied.

Upon receiving a share update, each participant updates their share by computing $(a_i, f(a_i) + h(a_i))$, where $s_i = (a_i, f(a_i))$ is their original share and $u_i = (a_i, h(a_i))$ is the received update. After computing the updated share each participant performs the Share Validation stage with their updated share and the updated commitment to ensure the integrity of their updated share. If share validation fails, the participant should delete the updated share and continue with their old share. If the new share is valid, the participant should delete the old share and store the updated share for future use.

Reconstruction. Reconstruction of s is as described in a classic Shamir threshold scheme. However, before recovering the secret, the participant performing the recovery first executes the Share Validate stage for each share. If a share is invalid, the Reconstruction stage requires the acquisition of a replacement share such that a set of t valid shares is produced.

4.9.3 Extended Protocol Description

Figure 4.5 summarizes the additional steps of our improved extended mode ceremony for Secret Preparation, Share Distribution$^+$, and Extended Reconstruction.

Secret Preparation. Secret Preparation now includes the generation of a separate integrity value for the ciphertext to be distributed in Share Distribution$^+$. This can be as simple as a collision-resistant hash of the ciphertext.

Share Distribution$^+$. Additional steps are added to the share distribution stage to enable ciphertext integrity in the Reconstruction stage. In addition to the shares, the above integrity value is distributed to each participant. Using a separate integrity primitive checked by the reconstructor during the Reconstruction stage protects integrity even against adversaries who know s. Each participant stores the integrity value along with their share in the selected safe storage mechanism.

Extended Reconstruction. Before decrypting the ciphertext as required, the participant performing the reconstruction checks their copy of the integrity value against the ciphertext.

4.9.4 Security and Limitations

We now apply our framework from Section 4.5 to assess the use of our defined improvements. We maintain the identical threat model and security goals as in our case studies from Section 4.8. We assume a high-powered adversary and desire the security goals of t-Separation of Privilege, Availability, Information Theoretic Security, Confidentiality, and Integrity.

Our improvements (summarized in Figures 4.4 and 4.5) guarantee Availability, Confidentiality, and Integrity of shares and the secret information for both the Base and Extended Modes of operation. Integrity is achieved due to the use of a proactive VSS scheme for share verification. We will now discuss each security goal in more depth and how it is achieved.

In Extended Mode, availability of the ciphertext is achieved via redundancy (defined in Step 2 of Share Distribution$^+$ in Figure 4.5). By distributing the ciphertext to each participant, n independent copies are made while only one is required to recover the secret $\mathcal{F}$. This approach falls within our assumed trust model, as giving a copy of the ciphertext to participants who are trusted with a share does not result in additional powers or capabilities for these participants. In this case, the safe storage mechanism becomes the set of participant devices which store the copy of the ciphertext, achieving availability via redundancy.

The 'Share Updates' stage allows for removing participants from participating in a future recovery. Removing participants enables the preservation of confidentiality with revocation.

For the Extended Mode, we require the dealer to distribute the ciphertext integrity value to each participant, as defined in Figure 4.5 (Step 2 of Share Distribution$^+$). Distributing

the ciphertext integrity value to each participant allows for any participant performing the Share Recovery stage to verify the integrity of the ciphertext while the integrity vector $\vec{C}$ is used to verify individual shares.

Requiring a Sufficiently Random Secret. Note that the commitment scheme for Base Mode requires a sufficiently random s. In the case that s is not sufficiently random, the entropy of s can be increased by moving to the Extended Mode by encrypting the low-entropy secret with a high-entropy key. Alternatively, the dealer can pad s with a sufficiently large number of random bits (e.g., 256 bits). As a final requirement on secrets and shares having sufficient entropy, we highlight the importance of generating secrets and shares on a machine with sufficient entropy sources to prevent amplifying an adversary's guessing attack [70].

4.9.5 Implementation

The techniques we employ in our implementation, specifically for proactive verified secret sharing, are derived from those of Herzberg et al. [92]. Our implementation differs slightly from their Share Update function, which requires *every* server in the threshold scheme to generate a new update value and distribute it to each other server in the system. In our protocol, any participant may generate an update and send it (noninteractively) to the other participants; the correctness of the update can be verified from the commitments. These other participants may be online or offline. If they are offline they will perform the update when they come back online. If a participant does not trust an update they can initiate another update. Finally, we do not require all participants to perform the update generation, or to be online. Therefore, our derived implementation (as initiated by any one participant) is noninteractive.

Our implementation is in Rust and uses curve25519-dalek [133] for group operations, which we make publicly available (https://crysp.uwaterloo.ca/software/vss/). Group operations are performed in Edwards form for speed and safety properties. The benefit of using Rust for our implementation is multi-language interoperability and memory safety. Furthermore, our changes can be integrated with implementations of threshold schemes in Rust such as RustySecrets [217] and Sunder [76].

4.10 Conclusion

Although the theoretical study of secret sharing protocols began decades ago, the use of secret sharing schemes in practice remains poorly defined. Interest in standardization of practical implementations of threshold cryptography is growing, including by NIST [25]. However, to enable practical use, researchers and practitioners must account for gaps in security that arise when moving from a theoretical setting to a real-world application. As a step towards practical secret sharing, we present a framework to facilitate the security analysis of threshold schemes based on Shamir secret sharing. We distinguish between operating in a base or an extended mode of operation, and through case studies, we demonstrate that variations in the ceremony of secret sharing schemes can lead to changes in the fundamental security properties provided to end users. Our framework can aid the design and analysis of future implementations of secret sharing by providing a more detailed ceremony definition and accounting for previously undefined assumptions about adversaries, user roles, and user actions or decisions within the scheme. Finally, we introduce and implement a secret-sharing protocol with improved security properties that can be directly integrated with existing Shamir secret sharing implementations.

Share Generation

1. Steps 1–4 as before in Figure 4.2
2. Device: Generates a commitment $\vec{C} = \langle \phi_0, \ldots, \phi_{t-1} \rangle$, where $\phi_0 = g^s$, and $\phi_j = g^{r_j}$ for $1 \leq j \leq t-1$.
3. Choice/Action: The dealer publishes $\vec{C}$ to a trusted public location.
4. Steps 5–6 from Figure 4.2

Share Distribution unchanged from Figure 4.2

Share Validation

1. Action: The participant fetches $\vec{C}$ from its trusted public location.
2. Device: Using $\phi_0, \ldots, \phi_{t-1}$ which constitute $\vec{C}$, the participant will then calculate ψ by evaluating $\prod_{j=0}^{t-1} \phi_j^{{a_i}^j}$.
3. Device: The participant validates her share if ψ is equal to $g^{f(a_i)}$.

Share Updates

1. Action: $\mathcal{U}$ executes the Share Generation stage, with unchanged values for t and n, to generate a new polynomial $h(x)$ where s=0. For each authorized participant holding a share $a_i, f(a_i)$, use $h(x)$ to generate a share as the update $u_i = (a_i, h(a_i))$.
2. Action: $\mathcal{U}$ publishes the updated commitments to the trusted public location.
3. Action: $\mathcal{U}$ distributes the update $u_i = (a_i, h(a_i))$ to the (authorized) participant with share $(a_i, f(a_i)$ for $1 \leq i \leq m$, where $m \leq n$.
4. Action/Device: Each participant will apply the share update u_i, to their share s_i to produce $s_i = (a_i, f(a_i) + h(a_i))$.

Reconstruction

1. Step 1 as before in Figure 4.2
2. Action/Device: P_r ensures the validity of each share using $\vec{C}$.
3. Step 2 from Figure 4.2

Figure 4.4: An Improved Base Mode Ceremony via Verifiable Secret Sharing (VSS) and proactive share updates

Secret Preparation

1. Steps 1–2 unchanged from Figure 4.3
2. Action: Generate integrity value for the ciphertext.
3. Choice: Select a safe storage mechanism for the ciphertext.
4. Action: Store the ciphertext.

Share Distribution$^+$

1. Steps 1–3 unchanged from Figure 4.2
2. Action: The dealer distributes ciphertext integrity value, along with a copy of the ciphertext, to P_i for $1 \leq i \leq n$.
3. Steps 4–6 from Figure 4.2
4. Action: Each participant stores their copy of the ciphertext integrity value with their share.

Extended Reconstruction

1. Action: Acquire ciphertext from selected location.
2. Action: Verify integrity value.
3. Device: Use recovered s to decrypt the ciphertext.

Figure 4.5: Improved Ceremony Framework for an Extended Mode of Operation

Chapter 5

Practicalities: Privacy and Attack Amplification

In this chapter we highlight how while there are many ways to develop technical privacy protections; changes in context or configurations can have unintended, and significant, impact on the privacy protections that are achieved.

5.1 Introduction

Collaborative machine learning techniques, such as federated learning [145], support training on datasets that have different owners, which can be advantageous for two or more data owners that want to generate an improved shared model while keeping their data separate [54, 145, 184, 248]. However, when training sets contain private or sensitive data, the machine learning process must be done in a way that further limits the risks the risk of privacy leakage [107, 156, 205]. This is particularly true with respect to data used for health research. The tight regulations that protect health data prevent the explicit sharing of data to train machine learning models [3, 164, 230]. Therefore, for health researchers hoping to learn from a collection of distributed private data, machine learning models that can be trained without risking data privacy have great appeal [63, 106, 107, 203, 209]. However, the risk of the machine learning process revealing sensitive data should not be underestimated when considering deployment in the real-world.

An adversary possessing a trained machine learning model can perform an inference attack to gain insights about the data in the training set [32, 207]. When the adversary's

inference goal is to determine whether a specified data point was in the training dataset, then they are performing a *membership inference* (MI) attack [104, 149, 207]. To understand some of the privacy implications of a membership inference attack involving health data, consider the following. Assume different research groups, all working on the same disease research (e.g., cancer, diabetes, etc.) used federated learning techniques on their distributed data. An adversary that can accurately determine whether an individual's data was used in training that model now knows that the individual has the disease being studied. Thus, we investigate further avenues to better understand the extent to which an adversary could succeed at that task.

While a number of MI attacks exist in the literature, they do not explicitly attempt to exploit aspects that can amplify MI attacks. Early investigation into why membership inference attacks work in stand-alone learning found a relationship between MI attacks and model over-fitting [226]. Additionally, in federated learning some authors observed that the number of participants affect inference attack success [149, 156]. However, past work has not recognized how existing MI attacks can be amplified when exploiting the specific features inherent to federated learning. In this work, we investigate the increased effectiveness of MI attacks given the ability of adversarial participants to observe model changes over time, use their own data to improve their attack effectiveness, and gain feedback on the success of their attack.

We highlight how features of federated learning amplify the effectiveness of MI attacks. This amplification comes both in improvements of attack accuracy as well as in the reduction of assumptions needed (e.g., expected loss) for the adversary to perform the attack. In short, federated learning has inherently more powerful adversaries than stand-alone machine learning.

In this work we define techniques for adversarial participants to amplify MI attacks. We empirically demonstrate how these techniques and different training conditions result in more accurate MI attacks. Thus, we illustrate the underestimation of MI attacks by showing the following:

- A MI attack from the stand-alone setting can be deployed per observed round of federated learning and achieve an amplified attack accuracy (increases of 15%).
- The total training data used, the local batch sizes, and the number of participants all impact MI attack accuracy in significant ways (including variations in attack accuracy of 23%).
- The additional data and training procedures inherently known by participating adversaries facilitates attacks that do not require any knowledge outside of what they

hold as participants (attack accuracy is maintained).

- Participating adversaries can learn the actual attack accuracy they achieve on their target model by executing the attack against their data.

Finally, we emphasize the implications for these results more broadly. Federated learning has previously been framed as a "privacy-preserving" solution in machine learning since it is not necessary to send datasets to other parties. However, as illustrated here, not sending datasets does not mean there are rigorous formal privacy claims that can be made by these systems.

5.2 Background

Federated learning. Federated learning aims to eliminate the need for data owners to share their datasets. In federated learning the set of participating data owners, each with their own datasets, each perform local training and each sending model updates to be aggregated. Generally, the updates are aggregated by a separate central aggregator that does not possess any data, who facilitates the training through performing computations [145, 205, 246]. Federated learning can be executed using model averaging [145], federated stochastic gradient descent [149], as well as model averaging with differential privacy [147].

To train a federated model, the central aggregator (or server) first initializes the appropriate parameters. Training then consists of global training rounds facilitated by the server and local training epochs performed by the clients (participants contributing data). After initialization, for each global training round, the server solicits updates from the clients who have trained locally on their own data. The server aggregates the collected client model updates by computing the average and returns the new global state to the clients. This process continues until either a preset number of rounds has completed or the model converges. Federated averaging (Algorithm 2 in the appendix) is one of the earliest federated learning algorithms, is widely used, and existing MI attacks focus on it [39, 97, 98, 125, 156, 247].

Membership inference. MI attacks encompass a category of attacks where an adversary aims to answer the decision problem as to whether a target x was a member of the training set used to produce a target machine learning model [97]. The term *membership experiments* encapsulates how researchers evaluate the risks associated with MI attacks

through a theoretical game. As part of the game, the adversary aims to discern whether a target data sample was a member of the training data. The game evaluates an adversary's success within the confines of what information is designated as being available to an attacker (the threat model). The threat model includes whether the adversaries may execute passive or active attacks, possess different adversarial knowledge (black-box, white-box), and any additional assumptions about attacker access and abilities.

Threat models. The two main threat models typically used in MI experiments are *black-box* and *white-box* access. While black-box access is more restrictive than white-box access in stand-alone machine learning, a participating adversary (or insider [225]) in federated learning inherently has a more complete view of the training process than either stand-alone access model. In federated learning, a participant trying to perform a MI attack against the other participants' training sets will have access to: their own datasets (which represent a subset of the global training data), their test data, architecture details, parameter details, and model access. That is, participants in federated learning have intrinsically more access than either white-box or black-box access attackers in the stand-alone setting, without requiring any additional assumptions as to how they get that access.

5.3 Related Work

Private machine learning is a technical solution that, in part, attempts to protect training data through the use of some selection of differential privacy [85], third parties [145, 167], and cryptographic computation [23, 225, 238]. Attacks that target machine learning models aim to thwart the privacy preserving designs; typically by either making modifications or inferences. While we focus on MI attacks [97], attacks on machine learning include property inference attacks, model inversion attacks, and others [22, 73, 83, 91, 94, 168, 240].

Membership inference attacks. In the case of MI attacks, an attackers' goal is to determine whether a target data element belongs to the training set that was used to train the machine learning model [207]. If, for example, belonging to a dataset is itself sensitive, MI attacks can be the most damaging attacks on the privacy of machine learning. Classifier-based MI attacks train additional classifier(s) to aid in the attack. These secondary classifiers may be trained in an imitation of the target model (a shadow model) over secondary dataset(s) [132, 149, 191, 192, 202, 206, 207, 213, 226]. Alternatively, MI attacks may rely on a threshold or heuristic to make their inferences, rather than training

any additional models. Threshold-based MI attacks may use the expected loss [243], prediction entropy [214], prediction confidence [192], and other threshold values [42, 101, 132, 215, 216].

Hui et al. [99] claim that threshold-based attacks, such as Yeom et al.'s [243] and the one in Salem et al.'s [192] lack the ability to effectively evaluate MI attacks. They suggest that there is insufficient access to have enough labeled input to identify a strong boundary between members and non-members. However, this is not the case for any adversarial participant in federated learning. Such participants always have a set of ground truth data that they can use in various ways to amplify their attack; including the ways we present in this work. Similarly, Irolla and Châtel [101] report that even if a MI attack has (relatively) high accuracy, attackers do not receive any indication as to the success of their attack on an individual target. That is, for a target x, they may know the attacks expected accuracy is $X\%$, but they do not have a way to discern whether the target x is classified correctly. While federated attackers still cannot compute the correctness of their classification for an individual target x, they are able to evaluate how well the attack is performing in the actual setting they are using it; through computing over their own members and non-members.

Attributing membership inference attack effectiveness. Since the first MI attacks in the literature, researchers have attempted to discern the features of machine learning that correspond to the success or failure of a MI attack [101, 115]. Past hypotheses attributed the success of MI attacks to the difference in confidence a model has for members versus non-members. One way this difference appears is in model over-fitting. Model over-fitting describes the case where a model has good accuracy on its training set (members), but poor accuracy on the test set (non-members, or new data) [223]. While MI attacks do perform well when targeting over-fitted models [101], over-fitting is not a necessary condition for a model to be vulnerable to MI attacks [42].

While Nasr et al. [156] attributed the attack success of Melis et al. [149] to the use of "unrealistic" training procedures, Liu et al. [131] found that MI attack success is heavily influenced by the complexity of the data being trained over. In part, they attributed this influence to the difficulty adversaries can have at acquiring a similarly complex dataset when the model is trained over complex data. Of course, this is not an issue in the federated setting where adversarial participants will hold an appropriate dataset since they contributed part of it. To summarize, understanding the privacy implications of MIs attacks requires accounting for all aspects of a training procedure. Further, as we will discuss later, the variance these factors create on attack accuracy could lead to adversaries that perform MI attacks in a new setting with higher or lower attack accuracy. A particularly

dangerous effect is if the adversaries can assess the accuracy of their attack within some reasonable bounds, as our adversarial participants do.

5.4 Attack Amplification

When evaluating privacy it is necessary to consider strong attackers. In federated learning, we assume the participating clients are adversarial, as they are powerful attackers in terms of the information they hold. Such adversarial participants, in addition to having access to the target model, are able to observe model states throughout the training process. Further, these participants hold their own data that they know is used in training because they contributed it.

In the following we define a series of techniques for amplifying MI attacks from a baseline. The attackers we present do not require additional information over what they would already possess as participants in the federated protocol. Our attacks are exclusively passive attacks (e.g., can be performed by an honest-but-curious adversary). That is, the adversary only observes information available to them and follows the federated learning protocol. They can perform additional computations outside of the protocol, but they do not deviate from the federated learning protocol and they do not provide incorrect or maliciously modified inputs.

Baseline. We classify all MI attacks that do not adapt their membership classifications by using information across training rounds as baseline attacks. This means both MI attacks targeted at stand-alone machine learning and MI attacks targeted at federated learning that do not adapt based on observations are considered baseline attacks. We use the original Yeom's attack [243] as our baseline (designated *vanilla*). This is a relatively straightforward MI attack. How it works, and why, is generally well understood. That being said, applying Yeom's attack's directly does not achieve a notably high attack accuracy when compared to more recent MI attacks from the literature. In Yeom's original threshold-based attack, the adversary is able to query the target model and learn the loss for the target element. Further, the adversary knows the average training loss of the model. The adversary makes the classification by evaluating $\mathcal{L}(x)$, where x is the target. If the loss for the target element is less than the expected training loss, then the target is evaluated to be a member, and a non-member otherwise. Thus, the expected training loss is used as the threshold at which the MI attack problem is decided.

Snapshots. When modifying a baseline attack for the federated setting, the first attribute we exploit is an adversarial participant's ability to observe changes over time (across training epochs). We define a *snapshot* as an observed model state. Each snapshot corresponds to a global model update sent by the central aggregator in federated learning to the adversarial participant who records it before using it to update their local model. After recording each snapshot, an adversary executes their baseline attack on the model's current state. Essentially, the adversary computes the base attack function (e.g., if the baseline is Yeom's threshold attack they computes the loss for the target), but for each training round they participate in. After performing the baseline attack for each snapshot, there is a classification of the target element x for each global training round observed (for each snapshot). To bring these inferences together and make a final classification, let n_e be the total number of training rounds (observed as snapshots). The function $Y_i(x)$ outputs the membership classification of the baseline attack for a target element x at snapshot i. Following a majority vote, the final formulation is if, $(\sum_{i=0}^{n_e} Y_i(x))/n_e) \geq 0.5$, than classify as a member. We refer to an attack that uses this mechanism as the *snapshot* attack.

Attack success feedback. The second attribute we exploit to amplify an adversarial participant's attack is their own dataset. As we will show in our evaluation, there are many features of federated learning that influence MI attack success. The implication of such variances is that an adversary performing an attack from the literature has no indication as to how successful their attack is. We show how adversarial participants are able to gain feedback on their attack success and how they can use it to tune their attack to be more effective.

Adversarial participants each hold their own dataset, which is a known subset of the models' training data. In other words, adversarial participants hold a set of known members (training set) and non-members (their test set). The adversary can use the same attack they are using for the target model, completely unchanged. The adversary then uses their own members (and non-members) as targets of the attack. After attacking their own data, since they know the ground truth as to which values are members and which are not, they can compute their local attack accuracy. We refer to this as a *self-attack*.

In addition to having an indicator as to how well the attack performs against their specific deployment, an adversary can also use the results of a self-attack to improve their success. An adversary can attempt to improve their attack, given the results of their self-attack, in one of two ways. They could either pick a new attack, or attempt to amplify their attack by excluding snapshot classifications with low self-attack accuracy. For each round the attack executes, if the self-attack accuracy is below $p\%$ (for any selected percent p),

then exclude the classifications from that round when computing the final classification via majority vote. We drop round classifications from the majority vote when the self-attack achieves 50% accuracy or less (thereby excluding classifications that perform worse than random guessing).

Algorithm 1 Client View Snapshot Distance MI Attack

```
/* Client executes: */
function ClientUpdate(θ)
    Save received model state a_i at epoch i
    for each target x do
        Compute ℓ_i = L(x)
    for each local epoch do
        for each batch B in client's data D do
            θ ← θ − η∇L(b; θ)
    return local updated θ
function Attack_YeomSnapshot({ℓ_i}_{i=1}^{n_e}, E[L], x)
    for each Snapshot loss ℓ_i do
        if ℓ_i < E[L] then
            set b_i = 1                                   // Guess Member
        else
            set b_i = 0                                   // Guess Non-Member
    n_e = total observed snapshots return (Σ_{i=0}^{n_e} Y_i(x))/n_e)
function Attack_ImitationSnapshot({ℓ_i}_{i=1}^{n_e}, x)
    Compute members and non-member ℓ's against snapshots based on client data
    for each Snapshot loss ℓ_i do
        Target x Difference Δ_x = ℓ_i − ℓ_{i+1}
        if Δ_x = 0 then
            set b_i = 1                                   // Guess Member
        else
            set b_i = 0                                   // Guess Non-Member
    n_e = total observed snapshots
    return (Σ_{i=0}^{n_e} Y_i(x))/n_e)
```

Reducing adversarial knowledge assumptions We demonstrate how an adversary can replace global model statistics used in MI attacks with their own approximations of the values. Additionally, we provide an alternative heuristic that incorporates the adversary's own datasets and epoch observations to execute their attacks.

An adversary can develop *imitations* of necessary values used in baseline attacks. For instance, Yeom's threshold attack assumes the adversary has access to the global expected loss. This is not something that the adversary will inherently know from the training. While it is possible this information is published as part of the model statistics, if the attack is weakened without the information, it would make sense to attempt to conceal it. Adversarial participants, however, would not be thwarted if they could not access the

global expected loss. Using their own members and the model, they can compute the expected loss across all their members and use that as their threshold value.

Our data driven heuristic attack is as follows. The adversary has a set $\mathcal{M}$ that they use in training (members), and a test set $\bar{\mathcal{M}}$ they do not use in training (non-members). Just as an adversary would compute $\mathcal{L}(x)$ for a target x, the adversary computes $\mathcal{L}(m)$ and $\mathcal{L}(\bar{m})$ for each $m \in \mathcal{M}$ and each $\bar{m} \in \bar{\mathcal{M}}$. These values are computed during each observed training epoch (snapshot) so they can be observed over time. To measure the change, the adversary computes $\Delta_{y,i} = |\mathcal{L}(y)_i - \mathcal{L}(y)_{i+1}|$, where i is the training epoch, and y is the element computed over (whether the target x, known member m, or known non-member $\bar{m}$. For the heuristic (MemDist, $Dist_i(x)$), if,

$$|\text{avg}(\Delta_{m,i}, \forall m \in \mathcal{M}) - \Delta_{x,i}| < |\text{avg}(\Delta_{\bar{m},i}, \forall m \in \bar{\mathcal{M}}) - \Delta_{x,i}|,$$

then classify x as a member in snapshot i, else, non-member. That is, if the targets' observed difference is closer to the average observed differences for members than for non-members at snapshot i, classify x as a member. Similar to executing Yeom's over snapshots, this produces a classification per round. In early rounds, if this attack is performing poorly (as evaluated against the adversary's own members), they exclude the inferences from that round. Similarly, if the attack is performing well against their own data (known members), they weight those rounds more heavily. We refer to this process as the *distance* attack.

Adversarial collusion. We identify two main features of federated learning that allow collusion to amplify an attack. First, a coalition of participating adversaries has access to more data. This allows them to have a better estimation of the behaviours of members versus non-members, and the ability to evaluate and tune their attack using their own members (e.g, computing $\text{avg}(\Delta_{m,i}, \forall m \in \mathcal{M})$ and tuning the attack per round). Additionally, consider the distance attack. If two or more participants want to collude with one another to estimate attack accuracy, one attacker can compute the thresholds for the distance attack (as usual), but then evaluate their attack using the members and non-members of the second attacker.

Second, an adversarial coalition is harder to deceive through excluding participants from some rounds. Including only a subset of participants can have at best a limited effect on reducing the model updates viewed by adversaries that have formed a coalition. Consider the following. Assume m of the n participants are included in each global training round and k of the n participants are adversaries. For such a coalition, of k participants, to miss seeing a global round (and thus miss a snapshot), there must be at least m honest participants, that is $m \leq n - k$. Then, even if $m \leq n - k$, the probability that only honest

participants are selected for a training round, and thus that a coalition will not receive the update for that training round, is $\frac{\binom{n-k}{m}}{\binom{n}{m}}$.

Finally, while we exclude coalitions of active attackers from our work, note that such participants could use carefully selected data to influence the results and improve upon their inferences (amplifying attacks such as from Nasr et al. [156]) via poisoning attacks [79]. Further, in the case of larger federated learning systems, where participants can self-enroll (such as users of predictive text keyboards), increasing the the coalition of poison attackers, can be done via a Sybil attack where the attacker enrolls large numbers of users as members of the adversarial coalition [58].

5.5 Empirical Evaluation

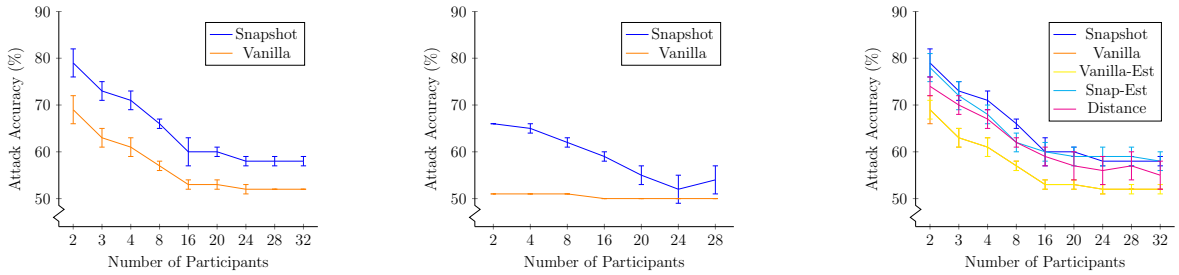
5.5.1 Experimental setup

Training target model. We use TensorFlow Federated [222] with federated-averaging to implement the federated learning models that are targeted by our attacks. All experiments were executed using a 2 TB RAM, 80 cores (Intel Xeon E7-8870 at 2.40 GHz) machine. See Appendix B.1 for all hyper-parameters and model configurations.

Datasets. Each participant in our federated learning receives a non-overlapping subset of the total dataset. We evaluate the attack amplification effects over two datasets. First, we use the EMNIST dataset that is included as a pre-processed federated version of the character and digit dataset in TensorFlow Federated. Within the pre-processed federated dataset, each client is assigned the data points that correspond to a unique writer. As a result, this limits the default size of each clients datasets to approximately 100 elements, unless contributions from different writers are combined. Second, we use CIFAR-10 following the configurations of McMahan et al. [145], which supports larger client datasets.

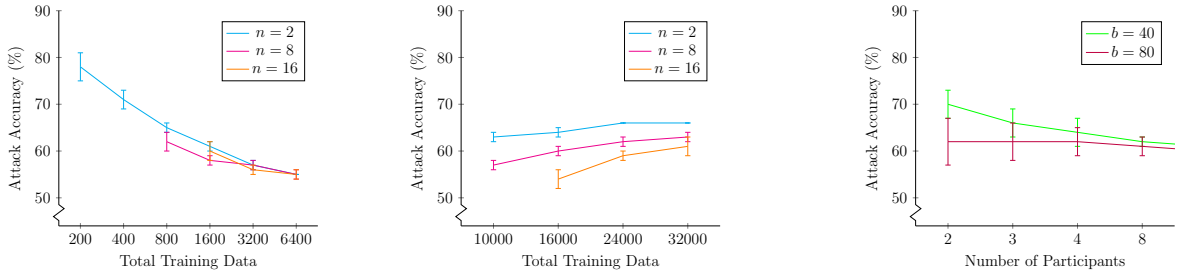
Attack setup. We designate Yeom's threshold-based MI attack (vanilla) as the baseline attack. We execute vanilla on the final model achieved after the federated training completes [243]. We evaluate each of: *vanilla attack*, *snapshot attack*, and *distance attack* while varying a series of federated learning features and demonstrating methods to tune attacks. For each attack configuration, we train ten target models, which are then each evaluated against the relevant configuration of features for the attack variants. We report standard

deviation and confidence intervals for each configuration (details in Appendix B.2). Our attack test set consists of 50% members and 50% non-members for all accuracy evaluations. Unless otherwise stated, target models are trained with the following configurations. For CIFAR-10, the default batch size is $b = 64$ and total training data is 24000. For EMNIST the default batch size is $b = 10$ and total training data is $n \cdot 100$, where n is the total participating clients. Our models are trained for 2-32 participants, focusing on the most vulnerable deployment scenarios relating to applications such as training models for health research over private medical data [106, 107, 149].



(a) EMNIST (b) CIFAR-10 (c) EMNIST

Figure 5.1: Comparison of algorithm amplification effect of snapshot over vanilla (baseline) for (a) EMNIST and (b) CIFAR-10, and (c) Estimate threshold attacks over EMNIST. The mean attack accuracy is shown and the bars indicate the 95% confidence interval of the mean.



(a) Snapshot, EMNIST (b) Snapshot, CIFAR (c) Snapshot, EMNIST

Figure 5.2: Attack accuracy degradation as data increases with fixed participants size.

5.5.2 Results

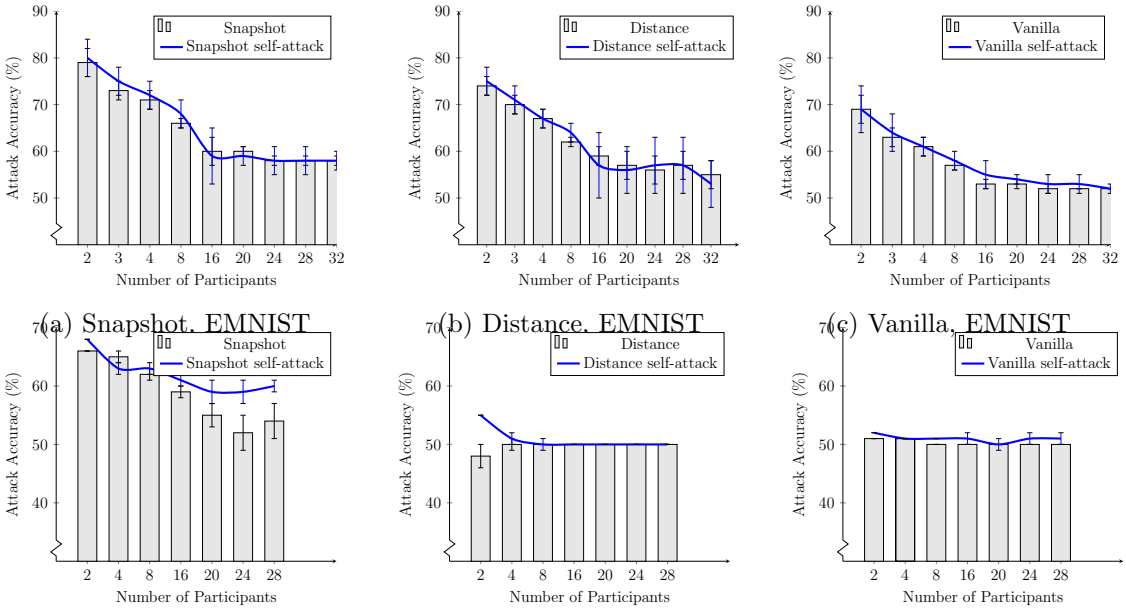
Algorithm amplification. We evaluate the vanilla attack and the snapshot attack for participating client sets of size two to 32. When the target models are trained using the EMNIST dataset, with a fixed amount of data per client, we found that snapshot attack

outperformed vanilla attack consistently, and with statistical significance. Snapshot attack outperformed vanilla by an average of 8% (when averaged across results for each participant set size), and achieves 10% amplification for the two, three, and four client setting (see Figure 5.1a).

The snapshot attack and vanilla attack against targets trained on CIFAR-10 are less successful overall then for the EMNIST trained models. However, the amplification from vanilla to snapshot persists (see Figure 5.1b). When the target models are trained using the CIFAR-10 dataset, with a fixed amount of data per client, snapshot attack outperforms vanilla attack with statistical significance until both attacks fall to 50% accuracy (at around $n = 16$ for this dataset). Against the models trained by CIFAR-10, snapshot outperformed vanilla by 15% for the two client setting. This means that even in a setting where an attack is otherwise believed to do no better than random guessing (50% accuracy) the attack is amplified to a better inference. The adversarial participants even maintain amplification when the global expected loss is not published. The adversaries can either use the average member loss instead (for Snapshot-Est and Vanilla-Est) or use our heuristic method and achieve similar attack amplification as snapshot (see Figure 5.1c).

Training procedures and amplification The snapshot attack still appears to degrade as the number of participating clients increases, inline with past work. However, the degradation does not push the attack to ineffectiveness until a higher number of participants as an artifact of the amplification. For instance, snapshot does not fall below 60% accuracy until there are 16 participants, while vanilla falls below 60% after the total number of clients surpasses four (for EMNIST models, vanilla never goes above 50% accuracy for CIFAR-10, recall Figures 5.1a and 5.1b).

We observe that the degradation previously correlated with increased participants has a confounding variable, namely the increased quantity of training data overall. We find that there is a similar degradation in attack accuracy when n is fixed and the total training data increases as shown in Figure 5.2a for EMNIST. Let $|\mathcal{D}|$ be the size of the total training data. Consider, for $n = 2$, the attack accuracy when $|\mathcal{D}| = 200$ is $78 \pm 3\%$ and when $|\mathcal{D}| = 6400$ the attack accuracy is $55 \pm 0\%$. Interestingly, when we repeat the same experiment of fixing n and increasing training data for CIFAR-10, we observe the opposite effect as shown in Figure 5.2b. Consider $n = 16$ for CIFAR. When $|\mathcal{D}| = 16000$, the attack accuracy is $54 \pm 2\%$, but when $|\mathcal{D}| = 32000$, the attack accuracy is $61 \pm 2\%$. Furthermore, these differences in attack accuracy occur without any corresponding variation in testing accuracy nor training accuracy. Thus, there is an attack accuracy difference of 23% in EMNIST and 7% in CIFAR-10 corresponding to changing the total amount of training



(a) Snapshot, EMNIST (b) Distance, EMNIST (c) Vanilla, EMNIST

(d) Snapshot, CIFAR (e) Distance, CIFAR (f) Vanilla, CIFAR

Figure 5.3: Comparisons of actual attack accuracy to the estimate by the adversarial participants.

data. Thus, the total amount of training data contributed to the federated model can significantly alter attack accuracy, and in ways that are not predictable in advance of deployment. When comparing the attack accuracy for different batch sizes ($b = 40$ and $b = 80$ for EMNIST), we found that larger batch sizes are correlated with lower attack accuracy (see Figure 5.2c).

Attack success feedback. We evaluated the self-attack with participant sizes 2-16 and found that it provides a close approximation of the overall attack accuracy. In Figure 5.3 we show the actual attack accuracy and the self-attack for each of snapshot, distance and vanilla against EMNIST and CIFAR-10 trained models. The self-attack provides a good approximation of the actual attack accuracy for both EMNIST and CIFAR-10. The accuracy of self-attack remains close to the actual attack regardless of whether the performance is good (e.g., CIFAR snapshot) or poor (e.g., CIFAR distance). Recall our results from earlier where changes in training can drastically affect accuracy (e.g., EMNIST for $n = 2$ total data resulting in 27% accuracy variance). If an adversary is unable to gain this feedback they have no indication as to whether they achieve, for instance, 50% accuracy or 77%. The adversarial participants can use their self-attack accuracy, to distinguish when

their attack is performing no better than random guessing (e.g., 50%) and when their attack is performing significantly better (e.g., 70%)

Thus, we have shown that adversarial participants in federated learning can know how successful their attack is and can therefore use this feedback to improve their inferences; including by dropping classifications from poor rounds (in Figure 5.4). We also note that the adversarial participants, having recorded snapshots of the model states, can independently perform each attack (vanilla, distance, snapshot) after the model training is complete. That is, if the feedback (e.g., for CIFAR) shows the distance attack is performing poorly, the adversary could then simply compute the snapshot attack instead.

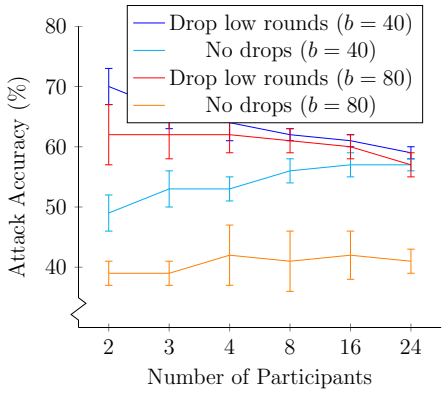


Figure 5.4: EMNIST Snapshot attack tuning.

5.6 Discussion

Our work illustrates how MI attacks are actually more successful than presumed when performed by adversarial participants in federated learning. We are, in effect, illustrating that proposals to apply federated learning to private data cannot make privacy guarantees given the range of avenues available to adversaries to perform an attack. We acknowledge that we only performed attacks on image datasets and we did not consider very large client sets of federated learning. However, we argue that participating clients contributing private data are likely to be low in number (e.g. recall the health care setting earlier with clients in the tens and not the 100s or 1000s). Therefore, we have chosen to focus on understanding the privacy risks for groups sized in the lower range.

We aim to prevent negative societal impacts by cautioning on how easily different configurations and computationally trivial calculations can enable an attacker in federated learning to make better inferences about private training data. We show that without requiring additional computational power or additional knowledge, participants in federated learning are able to make inferences about the training data, because they have a subset of it (their own training set) to compare to. Further, the power gained by adversarial participants because they hold their own dataset is so critical to any privacy claims for federated

learning because of how unavoidable it is. The adversarial participants need their data to contribute, so it cannot be restricted to prevent it from being an attack vector. They see model states over time because they help construct them, so these changes are not viably hidden. Researchers and practitioners should be aware of the weaknesses we have identified before developing or applying federated learning for real-world privacy sensitive applications.

5.7 Conclusion

Before machine learning models can reasonably be trained over private data, it must be ensured that the data can be protected. In this work we have shown that even otherwise small changes to how a model is trained (e.g., batch size) can have a notable impact on a MI attacks' accuracy. Within federated learning, adversaries can both evaluate an attack across training rounds and perform a self-attack to learn how well the attack is performing in the actual setting. This allows them to tune their attacks to increase the accuracy and make more confident MI attacks. While we only evaluated variations of threshold attacks, the intuition behind the attack amplification will still stand for other attacks (e.g., sample-to-user-level MIAs against NLP[42]); specifically evaluating attacks per round and checking attack accuracy on their own data (per round and globally). Further, the adversaries do not require additional assumptions with respect to their knowledge beyond that of what they already know as participants. Thus, since no additional assumptions are required to achieve the amplifications, it is much harder to detect or counteract. Therefore, we conclude that current MI attacks provide only a lower bound on attack accuracy given that the adversaries in federated learning are far more powerful than those in stand-alone learning.

Chapter 6

Communication of Privacy

The concept of data privacy, and correspondingly the regulation and protection of it, is relatively new within our society. Even when limited to modern times, the initialization of data protection laws is hundreds of years behind other regulations that are part of our day-to-day life. For example, consider the regulations on food products. The original Massachusetts "Act Against Selling Unwholesome Provisions" came about in 1784 [134]. Meanwhile, one of the first data protection laws is not enacted until 1970 in Germany when the Bundesdatenschutzgesetz (federal data protection act) is introduced [150]. Thus, insofar as laws are meant to reflect the norms of society, we need to determine how to communicate the nature of private computation to better understand what rules and regulations can be put in place to protect members of the populace while employing advancements such as private machine learning.

6.1 Introduction

As data access and collection have grown, so have companies' attempts to leverage that data, with regulations trailing far behind. Collaborations between companies increasingly involve data sharing and disclosure. For example, Mastercard sold transaction data to Google to track whether Google ran digital ads that led to a sale at a physical store (i.e., evaluating ad conversion) [21], raising privacy concerns for data subjects.

Within such modern data sharing practices, a *data subject* is an entity whose data is present in the data set, while a *data controller* is an entity holding a dataset they are contributing to some analysis. Data controllers who are not themselves the data subject

may have different privacy expectations or requirements compared to when the data subject themselves directly contributes their data. The data subject may not have understood their data could be shared or sold [66, 130, 180, 230].

Private computation, encompassing complex cryptographic techniques like private set intersection (**PSI**) [38, 175] and multi-party computation (**MPC**) [86, 241], allows companies to analyze data while maintaining data subjects' privacy in many cases. The cryptography literature emphasizes the value of private computation for cases where the data is especially sensitive (e.g., health or financial data) [225], among mutually suspicious entities [30, 50], or when there are less open trust boundaries [225].

For example, at its essence, PSI refers to a computation where two or more parties each hold a private data set, but wish to collectively compute the intersections of their sets. The intersection can then be shared with one or more of the participating parties. For example, two companies could determine which users they have in common without disclosing the identities of the users not in common. PSI, as with many other private computations, can be implemented using homomorphic encryption, differential privacy, or combinations of techniques that produce different guarantees and efficiencies. Specific privacy guarantees follow from specific mechanisms used in the implementation, such as statistical assumptions or computational hardness.

While private computation is often substantially more computationally expensive and more complex than its non-private analogue, there is an assumption that it is in some way *better*. For instance, it is presumed to be better for privacy that when PSI is used, data is only shared about clients the organizations have in common. To date, the degree to which users perceive private computation as better, or even feasible and plausible, has remained an open question. An unfortunate state given that users' are the entities whose data is used in such computations. Similarly, despite a flurry of recent work investigating users' expectations of differential privacy [27, 122, 123, 237] and attempting to improve communication about differential privacy [45, 51, 72, 112, 154], users' attitudes about, and expectations for, the broader range of techniques subsumed under private computation has remained open. The only user-centered work on private computation [6, 211] has investigated usability from an expert's, rather than an end user's, perspective.

To recap, when an organization considers deploying private computation, two key attributes must be addressed: (1) what privacy guarantees can actually be made to data subjects and (2) are those guarantees meaningful to the data subjects whose privacy they aim to protect? In this work, we investigate the second question through 22 semi-structured interviews.

Without knowing what data subjects understand and expect from private computation,

one cannot develop tools that empower them to make informed choices. Thus, in this work we ask and answer the following research questions (**RQs**):

- **RQ 1**: What do data subjects understand about private computation, and how can specific examples facilitate their understanding of the concept? *See Sections 6.5.1.*
- **RQ 2**: How is a data subjects' willingness to share their data impacted when informed of private computation's properties (protections and guarantees)? *See Sections 6.5.2–6.5.3.*
- **RQ 3**: How do data subjects perceive private computation's risks (e.g., inference attacks and beyond)? *See Section 6.5.4–6.5.5.*
- **RQ 4**: How are perceptions of companies influenced by the use of private computation? *See Section 6.5.6.*

In brief, we found the following implications for private computation in practice. First, data subjects are able to evaluate and understand the implications of private computation over their own data. Thus, neglecting to inform them of such practices is denying them autonomy over their own data. Second, while participants have an appreciation for the protections private computation can produce, they do not find these protections sufficient to overcome the need for i) consent and ii) transparency. That is, there are general details participants' that are factors in their evaluation of acceptability (Section 6.5.3) that companies should communicate. Third, participants are aware of unique high-risk threat models for themselves and others that private computation cannot guarantee protection against (Section 6.5.4). Therefore, failing to communicate the implications of private computation practices can open up unintended risks for users and clients.

6.2 Background on Private Computation

Private computation is the suite of techniques whose understanding by a broad range of users is this work's focus. To provide context for user-centered communications, including highlighting the types of guarantees private computation provides, this section provides technical background information. Notions of private computation revolve around two key aspects: *what* is being protected, and from *whom*. Technical privacy guarantees a set of protections given a series of assumptions are met. The assumptions can be about potential adversaries, system complexities, or statistics. When these guarantees are not in place, private information may leak.

A private computation executes a function over an input to produce an output such that there are limitations as to what can and cannot be inferred by an adversary, even if the adversary possesses some form of additional data. The function enforces the limitations through the use of mathematical protection mechanisms from cryptography (e.g., homomorphic encryption) or statistical guarantees (e.g., differential privacy), or some combination of techniques. Such computations may be between two or more parties, and they may involve trusted third parties. What is being protected within private computation typically falls under one of the following two classes:

Class 1: Private Data Set, Public Results. Consider a scenario where one or more parties have a (joint) data set and want to release an analysis of the data set. For example, the Census Bureau may wish to release statistics about the population of a certain region. Abstractly, their analysis y is a function f of the data set D, i.e., $y = f(D)$. The party performing the analysis can employ a protection measure like **differential privacy (DP)** [61], which ensures that a single record in the data set D has bounded impact on the analysis y. That is, the output distribution of y shifts by at most a factor determined by a privacy parameter specified by the analyst. By bounding the impact of a single record, the individual records in the dataset have a measure of protection against being revealed to those who access the results of the analysis. Thus, the analysis becomes a *private* version of the computation with protections that bound privacy risk.

The data set D may be distributed among several parties (e.g., D_1, D_2). For example, a government may be interested in the wages of its student population and thus wish to intersect tax filings with various universities' registration records. Here, the analysis y may be computed as a **secure (multi-party) computation (MPC)** [86, 241], which is a cryptographic protocol enabling the parties to compute the function $y = f(D_1, D_2, \ldots)$ while ensuring that no party i learns anything except y and D_i. While differential privacy was the protection mechanism in the aforementioned computation, computations may use both mechanisms. That is, differential privacy and secure computation are composable.

Class 2: Private Data Set, Public Subset. While the previous computations protected all individual data records while revealing the output of a computation, we now discuss a class of protection style that instead aims to publicly (or selectively) reveal a subset of the data. Consider a case where parties want to learn additional information about their data or information about a relationship between datasets they each hold individually. For example, assume Google holds a set of ad views on the Internet and Mastercard holds a set of credit card transactions [21]. Google may want to learn which

ad views led to credit card transactions, while Mastercard may want to learn which transactions were preceded by an online ad. Abstractly, given a common identifier in the data, the two parties could learn the intersection of their sets. The process of learning this intersection while protecting their respective datasets is known as **private set intersection (PSI)** [75]. Two or more parties can compute the intersection of their data without revealing data they possess outside of the intersection using private set intersection (PSI). In particular, PSI reveals no information about identifiers not in the other party's set, but fully reveals each identifier in common (which may be assumed to already be known). Differential privacy can be used on the datasets for additional privacy [87], and extended forms of PSI can compute a function over the intersection [175].

Attacks on Private Computation. So far, we have defined what private computation protects. However, given that some information is revealed intentionally as part of a private computation, there are some risks. Recall that we reveal an analysis y as a function of a data set D: $y = f(D)$. Given y, it is possible for an adversary to compute the inverse of function f and obtain a set of possible data set(s) D. This inverse can be computed when given only y, but the adversary may also have background knowledge in the form of a probability distribution over the possible data sets D, further restricting possible inputs and thus improving the attack.

Inference attacks, a subject of ongoing research, may pose significant privacy risks for subjects in the data set D. For statistical datasets, the risk of **de-annonymization attacks** or other information leakage can come via the execution of summation queries [136]. In the case of machine learning, attacks may use queries to the model and other attributes.

We give a few examples from machine learning where the output y (given to the adversary) is a publicly released machine learning model (e.g., a neural network), the outputs during a distributed learning process (e.g., federated learning [146]), or both. A **model inversion attack** [74, 95] computes the most likely input for one class of the model. For example, for a face recognition model this can be a picture of the recognized person. A **property inference attack** [81] computes a property of the records in the data set given a description of the property. For example, for a face recognition model this can be the ethnicity of the recognized person. A **membership inference attack** [208, 244] computes whether or not a given candidate was part of the data set D. For example, for a medical classification model, this can be whether or not a patient's record was included in the study.

Inference attacks are still feasible if the adversary cannot enumerate all possible data sets D, since they only need to estimate the most likely inference. Differentially private

protection mechanisms complicate inference attacks [244], but their theoretical analysis is complicated and error-prone [100].

6.3 Related Work

Communicating Differential Privacy and MPC. While some past work has investigated expectations and understanding of multi-party computation, it has been limited to stakeholders other than the data subjects. For example, Qin et al. focus on the usability of multi-party computation from more traditional functionality perspectives [179]. Similarly, Agrawal et al. investigated the perspectives of specialists such as industry professionals, researchers, designers, and policy makers [6]. They found that these specialist participants described private computation as a tool for enabling organizations to work with data. While these specialists acknowledged the importance of end users (data subjects), few prioritized end users' understanding of private computation, increasing the risk that private computation could be used for privacy theater [211].

The technical privacy mechanism that is differential privacy, and its implications for end users, has received a lot of attention from the HCI research community. Efforts have been made to explain differential privacy using a variety of techniques [45, 51, 72, 112, 154] and to evaluate whether differential privacy improves users' willingness to share their data [27, 122, 123, 237]. Within those efforts there are attempts to convey risk using visuals, risk notifications, and metaphors. In part, the complexity of some of these illustrations can be attributed to the "oddness" of differential privacy. Differential privacy provides guarantees in the form of "two neighboring datasets are indistinguishable within some probability", and understanding that guarantee requires first understanding the notion of neighboring datasets. However, while past work has done an excellent job at investigating differential privacy, it is too narrow to encompass the implications that correspond to the use of private computation. Private computation, as described in Section 6.2, encompasses all such computational efforts by organizations where there are protected inputs and revealed outputs, using some protection mechanism, such as, but not necessarily, differential privacy. Therefore, over the course of an interview, we employ what is essentially the process of self-explanation for learning [8, 40, 41]. Self-explanation helps learners adjust their understanding of a topic through examples and explaining concepts back to others. Essentially, it is an inductive, generative process of learning private computation rather than a prescriptive learning process.

Perceptions and Preferences. Previous work has frequently found users to be averse to their data being used by organizations [71, 111, 139, 140, 180, 204]. As we mentioned earlier, a motivator for the use of private computation is the assumption that it will counteract this aversion. Therefore, it is necessary to study users' awareness, understanding, and motivations of both technical tools and their implications for individual and societal privacy [11, 51, 160, 188, 220]. Information about individuals may be collected by employers, government entities, and friends. Which of these collectors originally receives the information is one component of the 'context' or social domain in which information is shared. Recent work, included as Chapter 3, found that when considering different contexts, represented by the number and type of participating companies, there is an observable influence on users' perceptions of the data sharing practices. Once the information is in a different context, whether via use or disclosure, it can no longer be assumed to meet privacy expectations [158]. In private computation, there is necessarily two or more organizations contributing their data. That is, private computation inherently results in a change of context that can influence participants perceptions and preferences.

Law and Policy. Legal notions of privacy are primarily framed in terms of protections for individuals from government and from corporations; with legal and financial penalties for non-compliance. The legal guarantees a company makes are typically encompassed within complex privacy policies [47, 159, 190]. These guarantees are enforced, as much as they are, by local data privacy laws. For example, Canada has PIPEDA, the Personal Information Protection and Electronic Documents Act [164], the United States has the Children's Online Privacy Protection Rule (COPPA) [229], the Health Insurance Portability and Accountability Act (HIPAA) [3] and the recent California Consumer Privacy Act (CCPA) [219], and members of the European Union have the General Data Protection Regulation (GDPR) [230].

Designers of private computation protocols have suggested that it can help "simplify the legal issues of information sharing" [176] and resolve privacy issues in various domains [49, 118, 174]. However, changing laws takes time while new technologies are in constant development, and thus these laws do not encompass current and future privacy requirements and expectations of private computation [135, 163]. Such legal regulations may impact individuals' perceptions of privacy and thus necessitate recruiting participants from the same local as one another.

6.4 Methods

We employ semi-structured interviews to allow us to follow up on participants' responses and allow participants to ask for clarification, as there has not been much prior work on users' understanding of, and expectations for, the broad range of private computation methods we consider. All participants received the same set of questions with the order shuffled as appropriate. Appendix C.2 contains the interview guide. We refined our procedure through pilot studies with five participants. Questions that participants found confusing were either removed or clarified. We do not include responses from the pilot study in our results. Ethics Board approval covered the design of the study, consent process, data analysis, and protection of the data collected.

6.4.1 Procedure

Selection of Interview Questions. To address our research questions we developed a selection of questions that when answered, collectively addressed the questions of our study. Our questions generate insight into both participants understanding of the systems as well as their perceptions of them. Further, we present a range of data leakage scenarios to understand how participants perceive such risks.

Formalities. Before starting an interview, we reminded participants that participation was voluntary, that audio was being recorded, and that they were encouraged to ask questions throughout. The interview proceeded through the seven parts detailed in the rest of this section: expectations, term awareness, private computation definition and example, computation scenario perceptions, inference attack perceptions, general perceptions, and a co-design activity.

Expectations and Term Awareness. The interview began with baseline questions to establish participants' existing perceptions. Participants were asked to "list some of the ways that you expect companies use data about you and others" and whether they had ever "come across" eight terms related to private computation that we presented in randomized order: "private computation," "encryption," "hashing," "multi-party computation," "differential privacy," "federated learning," "private machine learning," and "secure computation." Terms with which participants were familiar resulted in follow-up questions about where they had come across the term, what they thought its purpose was for companies and individuals, and a request to define the term in their own words.

Private Computation Definition. We then clarified "private computation" for participants by defining and comparing a non-private computation with a private computation. After participants had the opportunity to ask questions, they were asked to consider what they thought could be an example of "a computation where the result could be made public, but the inputs used to determine that result were sensitive and needed to stay private."

Computation Scenarios. As one of the key parts of our investigation, we gathered participants' perceptions of, and expectations for, private computation through discussing four scenarios in randomized order. We presented participants with a selection of scenarios in which private computation could be suitably applied to establish a baseline with which to understand their responses to the non-private computation version versus the private computation version we subsequently described to them. Each scenario consisted of an overall description of the goal of the computation, as well as two ways this goal could be achieved. One way used a straightforward approach involving non-private computation, while the other way employed private computation.

For each scenario, we asked participants how acceptable they found each way of achieving the goal, as well as why. Their explanations and reasoning helped us identify what factors most influence perceptions of (non-)private computation. We also asked participants what differences they perceived between the straightforward computation and private computation in that scenario, how feasible they considered the private computation to be, and how the company performing data analysis might explain the private computation to users.

We select four scenarios to correspond to real-world applications that are permissible under some conditions. Specifically, the four scenarios involved wage equity [44], ad conversion [21], contact discovery [52], and census data [2]. The *wage equity scenario* described an organization collecting salary data with the goal of generating a report on inequities. The *ad conversion scenario* described a credit card company and an online company comparing their data with the goal of determining if digital ads lead to sales in physical stores. The *contact discovery scenario* described a social media company with the goal of determining whether a new user had contacts that already use the app. Finally, the *census scenario* described a government body collecting a range of data with the goal of informing policies and resource management, as well as making results public. The interview guide in Appendix C.2 contains the full description of each scenario. These scenarios represent three different private computation settings. Ad conversion and contact discovery are settings where PSI can be deployed, wage equity efforts can use MPC, and census data can use privacy preserving query procedures.

Inference Attack Perceptions. We then presented participants with four descriptions corresponding to a type of inference attack. For each, we gave participants a series of examples of what specifically the company could learn, asking the participant to explain how acceptable they found that situation. For instance, in the case of a membership inference attack, we said, "One of the participating companies will additionally *be able to learn which specific records in the computed result correspond to you*." The membership inference case examples included the dataset being a set of dating app members, a set of frequent drug users, a set of low-income households, and a set of people with a specific health condition. For each example, participants were asked how acceptable it is if the organizations involved could determine they were a member of the example dataset, as well as to explain their reasoning. The other attacks corresponded to model inversion attacks, statistical inference attacks, and property inference attacks.

General Perceptions. At this point, participants had engaged with four private computation scenarios, as well as four types of inference attacks. To unite these ideas, we asked how the participants thought companies should be communicating how they used data (with and without private computation) and what the companies' responsibilities to their customers were.

Co-Design Activity. We concluded the interview with a co-design activity that built upon all topics participants engaged with throughout the study [210]. We asked participants to pretend they were working at an organization that hoped to use private computation and then consider how they would choose to explain private computation to their customers or clients. Participants were able to write, draw, verbally respond, or use whatever other means of communication they preferred. After providing their own explanation, participants were shown all previous participants' responses to the question and asked what they would add from their own to that explanation and what (if anything) they would remove from it until they arrived at their final version of the explanation.

6.4.2 Participant Recruitment

We recruited participants based in the USA via the Prolific crowdsourcing service[1] using a survey that included demographic information and when they could be available for a synchronous hour-long interview over a video call. We kept interviewing new participants

[1] https://www.prolific.co/

until reaching saturation (no longer finding new themes). We seemed to have reached saturation with just under 20 interviews, but we performed a few extra to be sure. Participants received $1.45 USD via Prolific for the initial scheduling survey (average time 4 minutes) and an additional $30 USD for participating in the interview. While most interviews lasted between 50 and 60 minutes, the shortest was 40 minutes and the longest 90 minutes. These times include debugging technical issues (e.g., fixing a microphone).

6.4.3 Participant Distribution

As detailed in Table 6.1, we interviewed 22 participants falling in the following age ranges: 18-24 (4 participants), 25-34 (8), 35-44 (6), 45-54 (2), and 55-64 (2). Among participants, 10 identified as a woman and 12 as a man, with no other gender identities being used. The participants fields of work span a broad range including politics, librarians, environmentalists, educators, insurance, health, music engineering, technology, personal assistants, chiropractics, and marketing. In terms of the highest level of education completed by the participants, the distribution is as follows: five participants had completed a graduate degree (Masters or PhD), eight completed a bachelors or associates degree, six completed some college but no degree, and three participants completed high school. Further, six participants reported that they "had an education in, or work in, the field of computer science, computer engineering, or IT" and of those one reported that they "had an education in, or work in, the field of cryptography." We note that the only restrictions on participation was age (18-65) and country of residence. The upper bound was due to requirements our Office of Research Ethics sets for including older participants. We chose not to exclude the participant who reported cryptography experience as during the interview it became clear their familiarity was overstated. Their responses did not differ from the participants without that reported background.

6.4.4 Incoming Knowledge and Expectations

Participants initial expectations for data usage could influence their perceptions of private computation. Thus, we started the study by asking participants what their expectations were and what terms they were familiar with. We present an overview of participants incoming knowledge and expectations in the following.

Expectations. Participants had expectations in terms of what data companies use (purchase history, demographics, search history, salary data, and user preferences), what com-

ID	Age	Gender	Education	Tech	Crypto
1	18-24	Woman	High School		
2	18-24	Woman	Bachelors		
3	35-44	Woman	High School		
4	45-54	Man	Bachelors		
5	25-34	Man	Grad School	✓	
6	55-64	Woman	Grad School		
7	18-24	Man	Some college	✓	
8	25-34	Woman	Bachelors		
9	25-34	Man	Bachelors		
10	25-34	Man	Grad School	✓	✓
11	45-54	Man	High School		
12	18-24	Man	Some college		
13	35-44	Woman	Bachelors		
14	25-34	Man	Some college	✓	
15	35-44	Man	Some college		
16	35-44	Man	Bachelors		
17	25-34	Man	Bachelors	✓	
18	35-44	Man	Grad School		
19	35-44	Woman	Some college		
20	55-64	Woman	Grad School		
21	25-34	Woman	Some college	✓	
22	25-34	Woman	Bachelors		

Table 6.1: Participants' demographics, including age range, gender, and highest education completed. Participants indicated whether they have an education or work experience in a tech-related field, as well as in cryptography in particular.

panies use the data for (financial gain, improving services, forging social connections, and personalization), and companies' responsibilities with respect to the data (anonymization, preventing re-identification). P8 emphasizes that despite being aware of companies' practices, they do not necessarily approve or agree with how companies use their data:

> "Even though I don't love that, I expect them to use it like for their marketing purposes [...] grow the bottom line of their business, to make money off of my data, and who I am as a person." (P8)

Participants have an expectation that companies are protecting the data entrusted to them, but P18 expressed concern that data usage practices may go beyond what they expect and be for reasons which they are not even aware of.

> "Of course, they may use it for other reasons, which I'm not even aware of." (P18)

Relevant Preexisting Knowledge As a proxy for identifying any preconceived notions participants may have about private computation, we showed participants a set of terms from the space (see Section 6.4.1). All participants expressed familiarity with the term encryption, with a few also being familiar with the term hashing. Familiarity with hashing was limited to those with a technical background who came across it as a data mapping strategy. All other terms either had no participants reporting familiarity or participants could not place the origins of the familiarity. In these cases, the participants guessed they either came across the phrase in terms and conditions or in news articles.

Source of Awareness. We surmise that the term encryption is thoroughly embedded in various facets of day-to-day life. Participants responded that they learned of encryption via leisure, education, employment, and when managing finances. However, encryption is not viewed as being particularly relevant to participants lives:

> "[It's] something that's used by techie people or politicians or people who are doing nefarious things, I don't think of encryption as guaranteeing things for individuals, like the lay public like myself." (P6)

Guarantees. On one side, participants expressed skepticism as to what tangible protections encryption can provide. Emphasis was made that there are "no guarantees" (P16) and that while it may provide some protections encryption does not make it impossible for malicious actors to access things. For those that are more optimistic of the protections, it was viewed as a means of making it more difficult for unauthorized persons to access the data.

Companies' Purpose. Some participants responded that encryption is used to provide the "illusion of security" (P8) while others thought encryption is used to provide "customers safety with their data" (P21). Ultimately, whether they had confidence in the protections or not, participants reported that company's use encryption to their own benefit; whether it is for protecting customer data, protecting proprietary information, gaining customers trust, or avoiding legal penalties.

Defining Encryption. In general, participants' definitions of encryption were not fully comprehensive, but they did show an understanding of encryption at a conceptual level. Essentially, participants highlighted that encryption changes the information it is applied to. These changes were referred to as "scrambling" (P20) and "masking or disguising" (P15) the information. Further, the changes are done with the goal of providing some security to the information such that it cannot be read by unintended recipients. These responses, regarding transformations, are most inline with what past work termed an **iterative** mental model of encryption [235].

6.4.5 Data Analysis

We recorded audio from each interview. We automatically transcribed the audio via speech-to-text software; afterwards, a member of the research team listened to each recording and corrected the automated transcriptions, as well as grouping responses by question and section of the interview. We analyzed this qualitative data using an inductive approach, allowing themes to emerge. Myself and my collaborator (Vasisht Duddu) extracted participant responses and then collaboratively clustered them according to similar sentiments and themes using the affinity mapping procedure [96, 114, 199]. Affinity mapping allows us to employ a team-based, collaborative approach to iteratively identify all aspects participants articulated when discussing their understanding of private computation, as well as private computation's implications. As part of the iterative affinity mapping process, after the two researchers formed initial clusters of participant quotes, they reviewed each quote within a theme to see what they had in common and discuss whether the quotes contained any points not encapsulated by others within that theme. Through iteration, we ensured that unique insights were not overshadowed by more prevalent ones. This process enabled us to capture the full range of attributes participants considered, as well as those that most commonly influenced their opinions. For example, consider the following. In terms of themes for responses to the acceptability of the ad conversion case, we identified: consent, privacy, benefits to the company, and low (perceived) sensitivity. Responses to contact discovery brought out themes of consent as well as benefits, limitations, perceived risk, and data minimization preferences. We reviewed emergent themes with respect to commonalities and differences across scenarios and questions to better understand participants priorities and concerns. These then became the structure of our findings.

6.4.6 Limitations

While we strived to ensure a diverse sample in many aspects, our participants represent a convenience sample and skew young (less than 20% were age 45+) and educated (69% had completed a bachelor's or graduate degree). Our participants are WEIRD (western, educated, industrialized, rich, and democratic), and we make no claims as to our results being representative of other population groups [198]. All of our scenarios are based upon typical cases in North America, where our participants live, and some examples may not be permitted by laws in other countries. Similarly, our scenarios may not cover data analysis tasks that might be both legal and common outside North America. Finally, as with other response-based studies, we acknowledge the potential for bias towards what participants perceive as socially desirable behaviour [186].

6.5 Results

The following results presentation centers around answering each of our research questions. That is, in terms of comprehension, we present the development of participants understanding of private computation from their first descriptions through to the final explanation they construct at the end of the interview. In terms of perceptions and influence on acceptability (RQ2 to RQ4) we evaluate any changes in perception between scenarios and participants' reported reasons for these changes. This enables us to better understand the influences with respect to phrasing versus actual impact as the interview format allowed participants to frame their reasoning in their own words. Thus, we identify themes participants use in their decision-making process when considering our data sharing scenarios, describe how private computation descriptions influence participants perceptions of the scenarios, and describe any impact private computation has on their expectations for companies' responsibilities.

6.5.1 Comprehension of Private Computation

We asked participants to produce a definition at three points throughout the interview; as an instance of a low-level assessment technique for evaluating learning and understanding of concepts [8, 40, 41]. We later repeat this technique after the participants have experienced the examples throughout the remainder of the study. We observe an increase in understanding via participants own explanations of private computation from their original

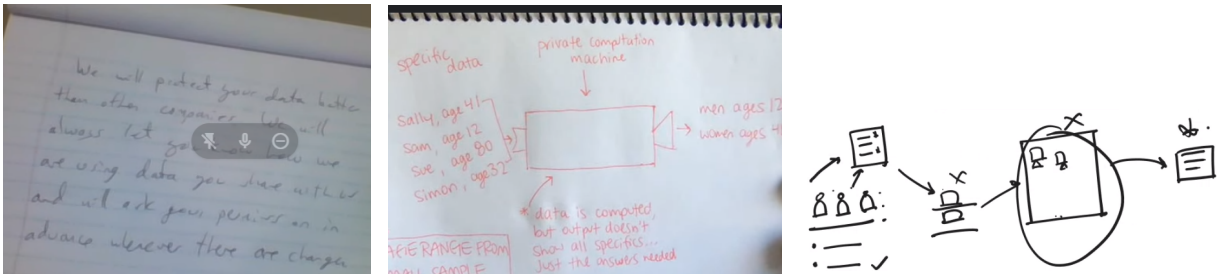


Figure 6.1: Participants used a range of mediums to convey private computation. Responses included written text, drawn images (digital and paper), and both verbal and typed responses. The above illustrations are from P6, P8, and P10, respectively.

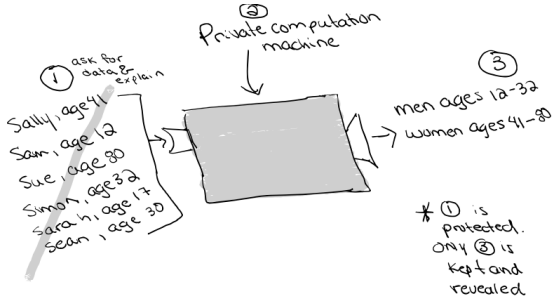
response at the start of the interview through to their final definition at the end of the interview.

First Attempts. They were first asked to provide their own description of private computation after they were shown a definition and asked to think of an example that could fit the definition. This definition occurred before participants were shown any of the scenarios included in the study. Participants struggled to provide an initial definition of private computation. Some participants were unable to come up with a definition. Of those that did provide a definition, they were generally brief, and typically overlapped with the initial definition they had been shown.

Participants did come up with several examples in response to the prompt for "an example of a computation where the result can be made public, but the numbers used to determine the result are sensitive and need to stay private". Not all participants came up with an example, some came up with more than one, and some participants changed their mind about their example (see Table C.1 in Appendix C.1 for the list of examples). The subject domains of the examples included salaries, research studies, and organizations profit data. The public outputs included aggregates, averages, company trends, and post-processed data. While not all of the examples were appropriate settings for private computation, the participants identified a number of cases that already exist. In particular, participants identified examples that corresponded to two of the scenarios we used later in the study; census data and wage equity.

Second Attempts. At the end of the study, the participants were asked a second time to explain private computation. At this point, they had seen all four private computation

Secure computation is a way that a company analyzes your data. The final analysis will be made public [at access location]. However, your specific data is protected and cannot be traced back to you nor can your specific data points be traced back to you. The analysis will be specifically [example], and this is being done because [purpose].



This is the information we're getting from you, but, rest assured, only Part Three will be shown. You can trust us to keep your information private. <If true> This information will only be used for this project and nothing else in the future.

Figure 6.2: Final explanation of private computation derived via input from the series of all interview participants.

scenarios and the cases corresponding to inferences attacks. For the second explanation they were informed that they could use any medium, including drawing a picture, verbal explanations, and writing. Participants' second attempt was overwhelmingly more successful than their first. Every participant provided a definition with their chosen medium varying (see Figure 6.1 for a selection of responses). Each definition was reasonably accurate, even if it was not all encompassing. Participants included in their descriptions what is being learned and what is being protected as important. Other aspects they suggested to include were how it will benefit the client and what the computation actually is. Further, in addition to their explanation, participants also identified attributes that they considered critical to quality explanations. Attributes participants emphasized are transparency and honesty. Participants also recommended including examples (especially as figures), summaries, and visually placing emphasis on critical points.

Final Explanation. Participants final definition is the one they derived after seeing the previous participants' final answer. Each participant was shown the explanation derived (by consensus) by the previous participants. They were then asked what they would add

or remove to the current explanation with consideration to their own initial response to the prompt. Earlier in the study participants made more dramatic changes, and they often incorporated large portions of their own explanations with smaller components of the current collective explanation. As the interview study progressed, participants made fewer and smaller changes, adding finesse as they identified attributes they considered valuable for an explanation being directed at the public.

When they made changes to the derived explanation, participants expressed the importance of clarity, accuracy, and conciseness. Participants emphasized that the value of being concise, but that it needs to be balanced with accuracy. For example, P17 noted that the original example would actually not protect the inputs:

> "The only thing I noticed is like, in this example, it's obvious the data is too small; that you can tell like the ages of specific men and women just because there's only two men and two women." (P17)

While it is the case that if "You add too much and you start losing it" (P16), without sufficient details, customers and clients could be confused or misled. Ultimately, participants made changes to improve clarity across steps in the illustration (consent, input, storage, output) and to add clarity on purpose.

For example, P6 found privacy failed to encapsulate what is being done and instead suggested using the term secure computation. They expressed concern that there is a dichotomy between privacy and using customer data such that private computation could never really represent what is being done: "If you're using my data, then there's no privacy [...] if there's privacy, then you're not using my data" (P6). This phrasing choice, which took private computation to secure computation was never changed by later participants. Further, other participants who noticed the explanation started with a different term, expressed support for the change and that "secure sounds better" (P8).

Also, in an effort to improve clarity, P8, introduced a visual example to the explanation. This illustration remained a core component of the final explanation, with other participants making small adjustments, but ultimately, they expressed an appreciation for the visual (P9, P10, P16-P18, P20).

The final derived explanation encompassed attributes participants emphasized throughout the interview process. Within the final answer, there is an explanation providing an overview of the concept, an example that walks you through the process (including permission to use the data being requested), the purpose, and a description of the guarantees being claimed. The final answer (after 22 participants) is shown in Figure 6.2.

As they constructed their explanations, participants did not focus on wanting to know the details of the mechanism used to achieve the guarantees. Participants trusted that the functionality was feasible without the details; leaving no need for complicated metaphors to prove it (Section 6.5.2). This eases the difficulty of communicating private computation practices in a way that is relevant, actionable, and understandable to the populace [194].

Based on the derived explanation, they did want to know the inputs, the outputs, the guarantees, and most of all the purpose of these computations. The components of the final derived explanation were: a description of the concept, what was being done and why, an illustrated example, and a brief explanation of the implications the computation could have for them. Further, these components are aligned with the themes that emerged when participants explained the acceptability of the four private computation scenarios. This consistency suggests these attributes are critical to members of the population being able to give informed consent to private computation. The remainder of this section revisits each of the components included in the final definition derived by participants; and provides insight into why these components were considered relevant by the participants.

6.5.2 General Impact of Private Computation

In evaluating our second research question, we found the following key points. Private computation may influence data subjects' willingness to share their data. However, this influence is not without its limitations. Participants expressed confidence in the ability to provide the guarantees described in the questions and that in many of the presented scenarios it made participants look more favorably upon the practice. However, as will be discussed in Section 6.5.3, private computation is not able to completely overcome factors previous work has found to matter to participants (e.g., purpose and consent).

Feasibility of Private Computation. In terms of feasibility, participants overwhelmingly considered the private computations in each scenario to be possible. Not only did participants think the scenarios were possible, but they thought such computations may already be happening (P12, P13 about census data). Participants did express concern, however, that companies may not be truthful about what they do with the information they collect (P22, contact discovery) and therefore thought it required some sort of enforcement. As participant emphasised, feasibility was not the critical factor, but instead that:

> "...it's you know, whether there are guards in place, it's do we have cops to to make sure that they're going to do what they're supposed to do." (P16)

Participants acknowledged that performing private computations could be more expensive (which was stated in scenarios where appropriate). When they considered the costs, participants included both the company's perspective and their personal views. Ultimately, while participants noted that companies may lose revenue by using such computations (P4 and P11), this was not considered to be an excuse to not protect their user's privacy. Participants even advocated that companies should spend more money on such projects to ensure that they are secure and safe (P2, P20, and P22).

Initial Perceptions of Scenarios Within our sample, participants generally perceived some scenario goals more positively than others. Specifically, the scenarios for wage equity and census data were generally positive, with responses clustering on the acceptable end of the scale (with few respondents considering these goals unacceptable). The scenarios for ad conversion and contact discovery, however, were viewed less positively. For both of these, the responses clustered on the unacceptable end of the scale. For instance, after they considered the contact discovery description, P14 responded that:

> "I want some privacy. I don't need a hundred percent. But, I'd like a little bit at least if that's not asking for too much" (P14).

Potential to Impact Acceptability For each scenario, participants view two descriptions, one corresponding to a private computation technique and one not, that could be used to achieve the organizations goals. The private computation descriptions used for both the ad conversion and the contact discovery scenarios see a positive change in acceptability. Wage equity has the most significant improvement with no participants reporting the private computation scenario to be unacceptable.

With respect to the private computation scenarios, the stipulations restricting the amount of data revealed and ensuring that companies cannot use the data for any other purposes are cited as improvements:

> "Even less of the data...data that is not relevant at all, they modify it to not make it available and I think that's, that's very thoughtful" (P9).

When considering the above attributes participants responded that "it feels a little bit more protected that way" (P12), "aligns a smidge more with my values" (P8), and "sounds like another layer of security" (P19). Overall, the descriptions corresponding to a private computation trend towards improving participants perceptions in terms of acceptability:

"they're not, you know, over exploiting what they're getting" (P22). The exception to the observed improvements with respect to acceptability is the scenario for census data; which actually has the opposite effect:

> "It feels like the second one's kind of saying the same thing. It just they're trying to make it sound a little bit better" (P19).

However, even for the more acceptable scenarios the improvement is not unconditional. Participants still express concerns for aspects that the private computations do not or cannot address. Ultimately learning something is the goal of any private computation, and that is not something that can be changed. As said by P7, "At the end of the day they're still like learning specific things about me." (P7)

Impact on Acceptability Due to Misconceptions. While some participants expressed exceptional insight into the risks and implications of private computation, others felt reassurance from its attributes. Unfortunately, not all of the attributes that gave reassurances actually provide the protections that participants expect. We identified two main concepts that participants find reassuring but are known to not provide the guarantees attributed to them. The first concept that provides false assurances is aggregation. For example, P6 described the protection from aggregation as:

> "When it's aggregated. It's lost. It cannot be disassembled. And private does not communicate that in any way shape, or form to me." (P6)

This confidence in averages and aggregation is unfortunately misplaced, as we know that there are a number of ways a malicious party could carefully select queries such that they can learn about an individual [136]. The idea that one can "blend into the crowd" via averages and aggregates and not experience additional risks is also observed in participants responses to the assorted inference attacks they were shown. That is, participants tended to find property inference attacks more acceptable than attacks that targeted an individual.

The second concept that provides false assurances is law, policy, and standardization. The assumption that the practices are "legal" or "industry standard" influenced acceptability. For example, P4 specifically stated that if the practice is not an industry standard then the acceptability would decrease. In the cases of P16, they concluded that if companies disclose such practices in their terms and conditions, it must be legal:

> "I don't know if in the real world, if this is legal to do, I would assume it's legal if they, if it's in their terms, right?" (P16)

However, while participants expressed confidence that the law protects against improper data sharing practices, this belief is not universal. Some, such as P11, stated that such practices do "not sound ethical [-] even if it's legal."

6.5.3 Bounded Impact of Private Computation

For each scenario, participants are asked how acceptable the scenario is and how companies should explain the private computation if they use it. Across scenarios, participants express a range of conditions that influence the acceptability. These conditions demonstrate limitations for private computation in terms of influencing data subjects' willingness to share their data.

Motives Matter. When responding to how acceptable they found a scenario to be, one of the conditions participants placed upon their answer was the goals and intentions of the company (P22) and whether the participants considered the reasons to be just and fair (P11). Goals that benefited society tended to shift their responses towards the acceptable end and goals that corresponded to corporate gain tended towards the unacceptable end. The scenarios for census data and wage equity were viewed as benefiting society. In the case of census data, participants went as far as to say: "it's like, crucial information gathering" (P8). Factors for participants when they viewed the census description included trust in the government, importance for society, and how such data is used:

> "And if the government is going to spend money, it may as well be based on some data rather than shooting from the hip." (P6)

Similarly, the wage equity description was considered to provide an important societal benefit that prioritized fairness and countered discrimination:

> "Wage equity should be a goal of a civilized society and companies aren't going to do that on their own. So third party organizations come in to try to ameliorate some of the inequity that the companies have within them" (P13)

Unlike when the organizations' goal is viewed to benefit society in some way, scenarios where the computation benefited the company were less positively received:

> "This is based on making more money, they're not considering the actual person involved" (P11).

In particular, the ad conversion scenario was seen as exploitative and unnecessary:

> Want to determine whether [...] their ads are effective? Well, you're still in business right? See, that for me, that's enough." (P16)

Some participants expressed that they understood why the company would want to perform such computations to determine if the money they spend on advertising was effective. Participants that expressed such understanding were still divided in that while some also thought it was fair, others thought companies should determine effectiveness without using additional personal data: "companies should have their own analytics [...] to figure out their own conversions" (P21).

Regulate the Restrictions. In the census case, there was actually an increase in the number of participants that consider the scenario to be unacceptable or completely unacceptable. Participants expressed concern both about the aspect of "any query" being permitted as well as about how query restrictions would be determined. Participants expressed concern that companies would exploit such restrictions such that "it's more like withholding information" (P18) and therefore they wanted to know "who is making the decisions regarding the information that's permitted" (P8).

Essentially, participants views were dependent on who makes the restrictions as well as what is restricted. P16 spoke about the importance of allowing the public to replicate results themselves whenever possible. They supported protecting individuals, but emphasized the importance of balancing protections with transparency:

> "If we're talking strictly numbers I lean towards all information available. There shouldn't be any math problem that that is is hidden." (P16)

This view was shared by other participants who also emphasized that the acceptability of such restrictions is highly dependent on the information in question:

> "...depending on what information is permitted, it might be good for somebody to know something that they're not permitting through the system, or it might be bad to let people know something." (P13)

Finally, some participants considered both descriptions to provide insufficient protections and desired additional restrictions (P5 and P10). These participants suggested a hybrid version of the descriptions to produce what they considered to be a more privacy preserving version. Specifically, to address their concerns, they suggested a query variant that only allows aggregate (or average) based queries while also preventing inferences beyond what is permitted.

Divulge the Details. Identifying what information individuals prioritized in their decision making is key to ensuring that information is communicated in the future. Participants mentioned a number of details for inclusion in explanations, and indicated such details are an influence on acceptability. In particular, participants who responded that a scenario was less than acceptable (e.g., neutral or unacceptable), emphasized that further information is required before the scenario could be acceptable. First and foremost, participants wanted to know that their data is being used:

> "That it's [the data is] being used. What's being done with it. The other company that is involved, that is Having access to it, and if it's going to be like ongoing or not." (P17)

Beyond knowing their data is being used, participants wanted to know how the data is being used. They wanted to know who is doing the computations and why they are being done. They wanted to know how long the data is being used for, how the data is protected (including the limits of those protections), and the implications for them if their data is used in these ways. For some participants, a failure to provide details or implement any of the protections the organization claims, are reasons to decline to participate in private computation. In other words, even when private computation is employed, participants care about appropriate flows of information [158]. Participants want to be allowed to judge if a flow is appropriate for themselves, and to do that, they require details with respect to the information flows.

Consent Above All. The details participants expect to be provided with are not just about the information. Rather, participants' desire to be informed is a means to an end, autonomy over their own data:

> "Every time your data is used in some kind of computation, you should be specifically alerted by the company; they shouldn't be able to do private computations [...] without you being aware of it." (P13)

A theme that emerged across all scenarios from the interview is consent and the importance of choice and communication to have meaningful consent. P17 summarized this notion as:

> "If they don't prompt you, then completely unacceptable. If they do prompt you then completely acceptable." (P17)

Participants, such as P1 and P16, both emphasized that consent is not a one-time thing. Companies need to be informing individuals periodically, or "every step of the way" (P16), about how their data is being used and ensure that they continue to consent:

> "When they sign up for the credit card and periodically, they should be reminded that all of their data is, you know, being sold to other companies." (P1)

In cases where participants may want to withdraw consent, the means to do so should be clear and accessible. Companies need to be "giving simple directions of, you know, where to go to opt out on the application" (P4). Such directions support individuals who change their mind about data use as well as those who did not understand or intend to agree:

> "Some kind of system where if a person finds out that they sign something that they really didn't understand, they can have a way to retract their permissions or whatever." (P13)

The final attribute participants emphasised as critical for consent is the use of clear and transparent communication. That companies need to be "proactive" and "not just rely on legal contracts to protect them". For instance, when informing individuals about how their data can be used, it should not be buried in terms and conditions nor obfuscated by legalese:

> "Be more upfront about how they're using our data instead of varying it in like really wordy terms and conditions in language that the average person like myself...like we can't understand it very well." (P1)

6.5.4 Risks for Unique Threat Models

We now discuss participants responses associated with the perceived risks and implications of private computation. In addition to the specific risks discussed at the end of the study (the inference attacks), participants highlighted what risks they perceived as possibilities in this space.

Participants question the potential implications of private computation and identify a number of risks associated with the contexts in which private computation could be applied. Both P13 and P19 identify risks associated with the goals of the scenarios, regardless of

the use of private computation. Individuals can be in situations where computing such connections could put someone's safety at risk. For instance, after considering the contact discovery scenario, P19 expressed concern that such connections could reveal someone's internet presence to an abusive ex or someone they have a restraining order on:

> "Through [...] common contacts that now he all of a sudden has a friend who has her information and now he has her information if through the tangled web, you could be able to find people [...] that's a growing problem." (P19)

Such risks are not things that can be resolved with a technical solution, such as private set intersection, but instead highlight the importance of informing users and gaining their consent, respecting their own risk assessments.

6.5.5 Inference Attacks and Acceptability

In terms of privacy leakage, the concern that organizations might make inferences from the limited information was brought up by P22 before any of the inference attacks were discussed.

> "If you're only giving like limited information, you might wonder if they're gonna acquire other personal information about you from that limited information." (P22)

Our participants also expressed concern that they "can't really figure out [...] the implication" (P6) of the computations or "how it could be exploited" (P15). As expressed by P22, the concern is that companies may request limited information, but try to gain additional information via some other means:

When presented with specific examples of information leakage, the perceived sensitivity or risk associated with inference attacks centered around two types being the most concerning. The first case being any instance where an individual is identified (e.g., membership inference attacks). The second instance was any instance where a group of people could be discriminated against (e.g, certain property inference attacks). Across all inference attack examples, the perceived sensitivity of the data is a factor for the acceptability. Location data, health data, sexual orientation, and religion are cases where the type of data is deemed to be more sensitive. Of particular concern were the cases that included health data. Participants, who were all located in the United States, expressed concern that their insurance company would get this information:

> "If that information then got shared with like my insurance company [they] would then decide to raise my rates because maybe I am at an increased risk for heart disease." (P1)

Among participants there was concern that the inferences made through the attacks could be used in malicious ways and to propagate bias and discrimination:

> "What this data is going to be used for, the state of it, should be used to to propel humanity forward. Not hold, not keep people back." (P16)

With respect to the inference attacks, some participants viewed all such attacks as unacceptable; since the companies were "not supposed to have that information period" (P6).

However, we did observe that inferences that target groups rather than individuals were less negatively viewed. Inferences about properties of groups are generally perceived to be somewhat more acceptable, however, this trend is conditional upon the specific property and the potential implication that property has for individuals and society. For instance, if the property could be used to "manipulate the populace" (P13), is "rude", or "discriminating" (P22), then participants state it would be less acceptable.

For conditional attacks, information leaks only occur probabilistically. However, this was not necessarily viewed as an improvement by participants:

> "It's based on what is what that record is, is in relation to even if it needs to be protected and it should be protected 100%." (P16)

Many found it unacceptable regardless of the percentages and stated that percentage was irrelevant. Of those that found a tipping point to neutral or acceptable they either tipped at 50%, 25%, or 1-2% chance the exact record would be learned.

6.5.6 Expectations for Responsibilities

An individual's ability to protect themselves is almost inconsequential without support. For example, after expounding on how a company's priority is their organization and financial gain, P6 expressed concern for how they are supposed to learn what they need to have data autonomy:

> "...how do I protect myself and who teaches me how to protect myself? Who's responsible for teaching me how to protect myself?" (P6)

Participants identified responsibilities for companies, government, and even themselves as individuals. Companies have the most responsibilities with respect to the law, protecting user data, and treating data with respect. The government's responsibility is to protect individuals via the creation and enforcement of policy.

Re-humanize Data. Participants expect companies to protect the data entrusted to them using the "best" security measures available to them as that data is not just some abstract input to compute over. Rather all of the data they hold corresponds to an "actual individual person with a name, a face" (P9). The data companies collect has been entrusted to them. Companies are expected to treat the data with respect and to be aware that the data is something important that they are responsible for. Treating data recklessly can have consequences for actual people:

> "I think the ultimate responsibility is to use it with caution. To protect people's privacy. It's up to the company to make sure they only share to the extent, the person allowed them to." (P9)

Furthermore, respecting the people who are represented by the data requires companies to exercise clear communication. Without transparency into data sharing practices people will continue to struggle to have autonomy over their data.

Proactive and Transparent Communication. When using customer data companies need to be upfront about their actions, but also provide greater granularity of control. For example, rather than a vague description for individuals to agree or disagree to, companies can be more specific:

> "You either agree or disagree and it doesn't really give much more information on what type of data is being used." (P12)

In addition to being specific, companies need to acquire explicit and ongoing permission for the collection and use of data. One participant even hypothesized that data sharing practices would be more positively received overall if there was not so much obfuscation and manipulation in the space:

> "A big social outcry [...] that could really be prevented if they were open from the very beginning. If people just knew, they wouldn't be so spooked by it." (P9)

Regulation and Enforcement. While some participants called for clearer regulations, some directly called for the practice of companies selling data to be made illegal:

> "They need to stop selling our information in general [...] passing that information to a company, I just I think it should be illegal." (P19)

However, in terms of law and regulation, participants tended to agree that companies have the responsibility to follow the law and the government has the responsibility to enforce the law, regardless of the use of private computation:

> "Health is a sensitive topic and and they're already legal protections for health information and so on. So...I don't see how why this edition of technology should should change those protections." (P16)

Participants made suggestions as to how the law can be enforced; specifically they suggested employing independent third parties. For instance, P21 suggested a third party could perform compliance checks and P1 suggested an independent entity to review points critical to consent. The independent party would perform a review to determine the best way to communicate to users about how their data is being used. They would also determine what information users need to make an informed choice about their data; such as a set of points everyone using the service should know about.

6.6 Discussion

Across our participant set, each individual demonstrated increased understanding (via explanatory evaluation) as well as communicated to the researchers factors related to private computation that influence their perceptions of the practices. Since the reasoning expressed by our participants included both traditional aspects for data sharing (purpose and transparency) as well as the technical guarantees (statistical-inference protection, property-inference protection, and membership inference protections) we detail in the following how researchers, developers, and policy makers can better communicate these aspects to the data subjects.

For Researchers. The improvements to acceptability generated by private computation were not universal. Even within the private computation examples participants were shown, the techniques used did not resolve all of their concerns. Participants identified implications of such computations, even proposing some alternative solutions, however, it remains the case that the implications of computations are not always clear. Thus, we recommend future work to develop more efficient communication techniques for private computation that includes details from our participants' final description (recall Figure 6.2). Further, future technical developments of private computation should consider the limitations of the protections they define and whether the trade-offs being made are improving data sharing practices. For example, when developing techniques that provide probabilistic privacy guarantees, researchers should include consideration for whether there exists cases such that there is justification for these guarantees.

For Policy Makers. Regulations related to private computation should account for how the descriptions of such practices influence data subjects' willingness to share their data, potentially more so than the actual guarantees. For example, our participants express confidence in the protections of aggregated computations and averages. In practice, this confidence is misplaced [136]. To ensure organizations with less than benevolent intentions do not use this confidence to propagate dark patterns [24], it is necessary to regulate how companies communicate practices such that they include implications and do not obfuscate them. We acknowledge that it is not necessarily possible, nor practical, to require companies to express all possible implications that could result from a computation they perform. However, whenever possible companies should be required to make explicit what protections are not possible as well as the limitations of the protections being employed.

For Companies. Participants expect the companies that use their data to treat it with respect. To treat the data as something important that has been entrusted to them and not something they have ownership over. While meaningful consent is a challenge to achieve, it goes a long way to fostering trust in an organization and willingness to provide data. Several attributes previously found to be relevant to individuals data privacy decisions are still very relevant within private computation. In particular, participants emphasized the importance of knowing the purpose or goal as well as a requirement that companies gain consent from their users before sharing their data. Even when using private computation, companies must communicate with the same level of transparency, including details related to how the computation is being used and what the company could potentially learn as a result of the computation. Communication should be transparent, accessible and clear

and the onus is on the companies to ensure they get informed consent. In short, a lot can be forgiven if permission is given.

For Consent. Similar to the study in Chapter 3, this work found participants placing a strong valuation on consent. To continue the discussion from that chapter, we emphasize that the valuation of consent does not inherently mean the desire to fill a collection of check-boxes and signatory acknowledgments. One aspect that may help address this desire is a better understanding of population privacy norms such that the standard practices of companies does not violate these norms. Continuing on this idea, if the norms are established and implemented it may inspire greater trust in companies treatment of data and the communication aspect will not be so repetitive. For instance, consider the following. We currently need to discover how best to communicate data sharing practices with the population for two reasons. First, to determine whether they approve of these practices and second to change practices. Going forward we could imagine a setting where default practices were informed by population preferences, determined via a "privacy census". Thus, people would only need to make these decisions every few years for their day-to-day practices and researchers can then focus on developing these population level communication practices. There may still be some extreme or unique cases that require at-usage consent and communication, but a "privacy census" approach could decrease the number of those and correspondingly decrease data-privacy fatigue and resignation to a loss of control often experienced by users (a sentiment expressed by participants in both studies included in this thesis).

6.7 Conclusion

While technical solutions are a powerful tool for protecting data, such protections do not directly correspond to personal privacy protections. The data being protected in these scenarios is not just an abstract concept, but instead is a placeholder for individuals with real lives and all the complexities that entails for their threat models. As a community, security and privacy researchers, data collectors, and policy makers need to remember that the protections provided by protocols and constructions do not and cannot encompass the full range of risks experienced by individuals in society. Technical privacy solutions must be conscious of the space which they may be deployed in and not guarantee that which cannot be delivered. This does not mean that such solutions do not add value, but that value must not be overstated. We must not forget that the data we speak of so abstractly is very concrete for the people whose lives generated it.

Chapter 7

Conclusion

There are personal and societal implications of private computation. These implications are controlled, in part, by the many stakeholders and decision makers involved in any practical deployment of private computation. For instance, there are developers, policy makers, researchers, data subjects, and data controllers (companies). Broadly speaking, this thesis has shown tools and techniques that make it easier to design meaningful privacy in machine learning such that it is clear and usable by all relevant parties. The parties considered go beyond those who own the data or organize the training of the model to include end-users who are the subjects of the data being calculated over.

I have shown that end-users may be more willing to share their data when informed of private computation, but not unconditionally (Chapter 6). That is, participants still found the context, such as the type of data and what was being learned, as critical to their willingness to share their data. More significantly, this study showed that members of the general population are successful at reasoning about private computation as it applies to them when presented with the salient information. Participants expressed what aspects met their needs as well as expressed concerns for unique risks that could apply to them regardless of the guarantees.

I have shown that attacks can be made more effective depending on the training settings in otherwise unexpected ways (Chapter 5). Not evaluating these attributes in a rigorous way creates a misunderstanding of the potential for privacy leakage or unnecessarily weaker deployment designs, as I have shown for comparatively simpler computations (Chapter 4). That is, based on the work of Chapter 4, I found that excluding the human steps from the protocol lead to dangerous gaps; and very different privacy guarantees. Therefore, when developing private computation protocols, including private machine learning, that are

intended for real world applications, I designed a series of studies that focused on defining end-users, who are data subjects', privacy preferences and needs.

I was also able to show that "types" of data sharing (number of parties, reciprocity, etc.) have different levels of perceived acceptability (Chapter 3). Further, variations in acceptability also exist between these types for controls such as data retention time and the purpose of the data sharing; all of which are controls that exist in designs of private machine learning. The implications of my results include showing that privacy policies that describe sharing with "trusted partners" or "trusted third-parties" do not provide the details relevant to individuals' privacy decisions. While long term it is likely to be beneficial for laws and regulations to account for these data practices and require informational granularity, such changes take time. More immediately, the implications of this research can be presented directly to organizations interested in using private computation and to fellow researchers designing novel protocols for these applications.

Future Work. While there is no shortage of future work needed within the topic considered by this thesis, I have chosen to highlight one for each of "perceptions" and "practicalities". The selected examples of future work, are perhaps the next most important steps for investigation within this space.

First, in terms of perceptions and communications, the next step is to design studies to evaluate the generalizability of approaches to communicating privacy-preserving machine learning and private computation more generally. Specifically, whether providing a description of the implications and impacts of the system rather than the procedures used can be generalised for a range of computational designs with descriptions for each "type". This will include the development of usable interfaces for accessible systems that make it clear where to get more info, the implications, how to opt-in or out; to essentially choose and have that choice happen first and not as a requisite for unrelated functionalities or services.

Second, in terms of practicalities, it is important to do an in-depth evaluation of the space to allow the design of a framework for what attributes of a private computation will impact its privacy outcomes as well as impact the populaces' perceptions of them. This will facilitate the design of novel protocols informed by the privacy needs and expectations identified. These protocols can be designed to meet both organizations' efficiency and functionality needs as well as the needs of the data subjects. Such a system will require a study with inputs from companies and organizations interested in using machine learning as well as a technical systematization of current abilities and limitations within the literature.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep Learning with Differential Privacy". In: *the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Vienna, Austria: ACM, 2016, pp. 308–318.

[2] John M Abowd. "The US Census Bureau Adopts Differential Privacy". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, New York: ACM, 2018, pp. 2867–2867.

[3] Accountability Act. "Health Insurance Portability and Accountability Act of 1996". In: *Public law* 104 (1996), p. 191.

[4] Alessandro Acquisti. "Privacy in Electronic Commerce and the Economics of Immediate Gratification". In: *Proceedings of the 5th ACM Conference on Electronic Commerce*. 2004, pp. 21–29.

[5] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online". In: *ACM Computing Surveys* 50.3 (2017), pp. 1–41.

[6] Nitin Agrawal, Reuben Binns, Max Van Kleek, Kim Laine, and Nigel Shadbolt. "Exploring design and governance challenges in the development of privacy-preserving computation". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–13.

[7] Irwin Altman. "The environment and social behavior: privacy, personal space, territory, and crowding." In: (1975).

[8] Thomas A Angelo and K Patricia Cross. *Classroom Assessment Techniques*. Ann Arbor, Michigan: Jossey Bass Wiley, 2012.

[9] Annie I Antón, Julia B Earp, and Jessica D Young. "How Internet Users' Privacy Concerns Have Evolved Since 2002". In: *IEEE Security & Privacy* 8.1 (2010), pp. 21–27.

[10] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. "Discovering Smart Home Internet of Things Privacy Norms Using Contextual Integrity". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.2 (2018), pp. 1–23.

[11] Erinn Atwater, Cecylia Bocovich, Urs Hengartner, Ed Lank, and Ian Goldberg. "Leading Johnny to water: Designing for usability and trust". In: *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*. 2015, pp. 69–88.

[12] Erinn Atwater and Ian Goldberg. "Shatter Secrets: Using Secret Sharing to Cross Borders with Encrypted Devices". In: *Cambridge International Workshop on Security Protocols*. Springer. 2018, pp. 289–294.

[13] Erinn Atwater and Urs Hengartner. "Shatter: Using Threshold Cryptography to Protect Single Users with Multiple Devices". In: *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM. 2016, pp. 91–102.

[14] Naveen Farag Awad and M. S. Krishnan. "The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization". In: *MIS Quarterly* 30.1 (2006), pp. 13–28. ISSN: 02767783. URL: http://www.jstor.org/stable/25148715.

[15] James Ball. *Unredacted US embassy cables available online after WikiLeaks breach*. https://www.theguardian.com/world/2011/sep/01/unredacted-us-embassy-cables-online. Accessed 2019-05-29.

[16] James Ball. *WikiLeaks publishes full cache of unredacted cables*. https://www.theguardian.com/media/2011/sep/02/wikileaks-publishes-cache-unredacted-cables. Accessed 2019-05-29.

[17] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. "Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from privacy Policy Text". In: *Proceedings of The Web Conference 2020*. 2020, pp. 1943–1954.

[18] Susanne Barth, Menno DT de Jong, Marianne Junger, Pieter H Hartel, and Janina C Roppelt. "Putting the Privacy Paradox to the Test: Online Privacy and Security Behaviors Among Users with Technical Knowledge, Privacy Awareness, and Financial Resources". In: *Telematics and Informatics* 41 (2019), pp. 55–69.

[19] Elana Beiser. "Record Number of Journalists Jailed as Turkey, China, Egypt Pay Scant Price for Repression". In: *Committee to Protect Journalists* (Dec. 13, 2017). URL: https://cpj.org/reports/2017/12/journalists-prison-jail-record-number-turkey-china-egypt.php (visited on 12/05/2018).

[20] Alastair R Beresford, Dorothea Kübler, and Sören Preibusch. "Unwillingness to Pay for Privacy: A Field Experiment". In: *Economics letters* 117.1 (2012), pp. 25–27.

[21] Mark Bergen and Jennifer Surane. *Google and Mastercard Cut a Secret Ad Deal to Track Retail Sales*. Online. https://www.bloomberg.com/news/articles/2018-08-30/google-and-mastercard-cut-a-secret-ad-deal-to-track-retail-sales. 2018.

[22] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. "Analyzing Federated Learning through an Adversarial Lens". In: *International Conference on Machine Learning*. Long Beach, California, USA: PMLR, 2019, pp. 634–643.

[23] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. "Practical Secure Aggregation for Privacy-Preserving Machine Learning". In: *the 2017 ACM SIGSAC Conference on Computer and Communications Security*. Dallas, TX, USA: ACM, 2017, pp. 1175–1191.

[24] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. "Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns." In: *Proceedings on Privacy Enhancing Technologies* 2016.4 (2016), pp. 237–254.

[25] Luís T.A.N. Brandão, Nicky Mouha, and Apostol Vassilev. *NISTIR 8214 Threshold Schemes for Cryptographic Primitives*. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8214.pdf. Accessed 2019-05-24. 2019.

[26] Brooke Bullek, Stephanie Garboski, Darakhshan J Mir, and Evan M Peck. "Towards understanding differential privacy: When do people trust randomized response technique?" In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 3833–3837.

[27] Brooke Bullek, Stephanie Garboski, Darakhshan J. Mir, and Evan M. Peck. "Towards Understanding Differential Privacy: When Do People Trust Randomized Response Technique?" In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, 3833–3837. ISBN: 9781450346559. DOI: 10.1145/3025453.3025698. URL: https://doi.org/10.1145/3025453.3025698.

[28] Eric W. Burger, Michael D. Goodman, Panos Kampanakis, and Kevin A. Zhu. "Taxonomy Model for Cyber Threat Intelligence Information Exchange Technologies". In: *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security*. Scottsdale, Arizona, USA: ACM, 2014, 51–60. ISBN: 9781450331517. DOI: 10.1145/2663876.2663883. URL: https://doi.org/10.1145/2663876.2663883.

[29] Jon Callas, Lutz Donnerhacke, Hal Finney, David Shaw, and Rodney Thayer. *OpenPGP Message Format*. https://tools.ietf.org/html/rfc4880. ID. 2007.

[30] Jan Camenisch and Gregory M Zaverucha. "Private Intersection of Certified Sets". In: *International Conference on Financial Cryptography and Data Security*. Berlin Heidelberg: Springer, 2009, pp. 108–127.

[31] Ran Canetti. "Universally composable security: A new paradigm for cryptographic protocols". In: *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*. IEEE. 2001, pp. 136–145.

[32] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. "The secret sharer: Evaluating and testing unintended memorization in neural networks". In: *28th USENIX Security Symposium (USENIX Security 19)*. 2019, pp. 267–284.

[33] Marcelo Carlomagno Carlos, Jean Everson Martina, Geraint Price, and Ricardo Felipe Custódio. "A Proposed Framework for Analysing Security Ceremonies". In: *SECRYPT*. 2012, pp. 440–445.

[34] Marcelo Carlomagno Carlos, Jean Everson Martina, Geraint Price, and Ricardo Felipe Custódio. "An updated threat model for security ceremonies". In: *Proceedings of the 28th annual ACM symposium on applied computing*. ACM. 2013, pp. 1836–1843.

[35] Microsoft News Center. *Microsoft and Providence St. Joseph Health Announce Strategic Alliance to Accelerate the Future of Care Delivery*. Online. https://news.microsoft.com/2019/07/08/microsoft-and-providence-st-joseph-health-announce-\strategic-alliance-to-accelerate-the-future-of-care-delivery/. 2019.

[36] Microsoft News Center. *Microsoft to Acquire LinkedIn*. Microsoft News Center. https://news.microsoft.com/2016/06/13/microsoft-to-acquire-linkedin/. 2016.

[37] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. "Differentially Private Empirical Risk Minimization". In: *Journal of Machine Learning Research* 12.Mar (2011), pp. 1069–1109.

[38] Hao Chen, Zhicong Huang, Kim Laine, and Peter Rindal. "Labeled PSI from Fully Homomorphic Encryption with Malicious Security". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM. New York, New York, USA: ACM, 2018, pp. 1223–1237.

[39] Jiale Chen, Jiale Zhang, Yanchao Zhao, Hao Han, Kun Zhu, and Bing Chen. "Beyond model-level membership privacy leakage: an adversarial approach in federated learning". In: *2020 29th International Conference on Computer Communications and Networks (ICCCN)*. IEEE. 2020, pp. 1–9.

[40] Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. "Self-explanations: How Students Study and Use Examples in Learning to Solve Problems". In: *Cognitive science* 13.2 (1989), pp. 145–182.

[41] Jennifer L Chiu and Michelene TH Chi. "Supporting Self-Explanation in the Classroom". In: *Applying science of learning in education: Infusing psychological science into the curriculum* (2014), pp. 91–103.

[42] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. "Label-only membership inference attacks". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1964–1974.

[43] Evgenia Christoforou, Pinar Barlas, and Jahna Otterbacher. "It's About Time: A View of Crowdsourced Data Before and During the Pandemic". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: ACM, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445317. URL: https://doi.org/10.1145/3411764.3445317.

[44] City of Boston. *Boston: Closing the Wage Gap*. https://www.cityofboston.gov-/images_documents/Boston_Closing%20the%20Wage%20Gap_Interventions%20-Report_tcm3-41353.pdf. Accessed 2022-09-23. 2013. (Visited on 09/2022).

[45] Amy Corman, Rachel Canaway, Chris Culnane, and Vanessa Teague. "Public Comprehension of Privacy Protections Applied to Health Data Shared for Research: An Australian Cross-Sectional Study". In: *International Journal of Medical Informatics* 167 (2022), p. 104859. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2022.104859. URL: https://www.sciencedirect.com/science/article/pii/S1386505622001733.

[46] Lorrie Faith Cranor. "Can Users Control Online Behavioral Advertising Effectively?" In: *IEEE Security & Privacy* 10.2 (2012), pp. 93–96.

[47] Lorrie Faith Cranor. "Necessary but Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice". In: *J. on Telecomm. & High Tech. L.* 10 (2012), p. 273.

[48] Lorrie Faith Cranor, Joseph Reagle, and Mark S Ackerman. "Beyond Concern: Understanding Net Users' Attitudes About Online Privacy". In: *The Internet upheaval: raising questions, seeking answers in communications policy* (2000), pp. 47–70.

[49] Emiliano De Cristofaro, Jihye Kim, and Gene Tsudik. "Linear-Complexity Private Set Intersection Protocols Secure in Malicious Model". In: *International Conference on the Theory and Application of Cryptology and Information Security*. Cham: Springer, 2010, pp. 213–231.

[50] Emiliano De Cristofaro and Gene Tsudik. "Practical Private Set Intersection Protocols with Linear Complexity". In: *International Conference on Financial Cryptography and Data Security*. Berlin Heidelberg: Springer, 2010, pp. 143–159.

[51] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. ""I Need a Better Description": An Investigation Into User Expectations For Differential Privacy". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. CCS '21. Virtual Event, Republic of Korea: Association for Computing Machinery, 2021, 3037–3052. ISBN: 9781450384544. DOI: 10.1145/3460120.3485252. URL: https://doi.org/10.1145/3460120.3485252.

[52] Emiliano De Cristofaro, Mark Manulis, and Bertram Poettering. "Private Discovery of Common Social Contacts". In: *International Journal of Information Security* 12.1 (2013), pp. 49–65.

[53] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata". In: *Science* 347.6221 (2015), pp. 536–539.

[54] Jeffrey Dean et al. "Large Scale Distributed Deep Networks". In: *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1223–1231. URL: http://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks.pdf.

[55] Department of Homeland Affairs, Australian Government. *The Assistance and Access Act 2018*. https://www.homeaffairs.gov.au/about-us/our-portfolios/national-security/lawful-access-telecommunications/data-encryption. Accessed 2019-05-20.

[56] Rachna Dhamija, J Doug Tygar, and Marti Hearst. "Why phishing works". In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM. 2006, pp. 581–590.

[57] Tobias Dienlin and Sabine Trepte. “Is the Privacy Paradox a Relic of the Past? An In-Depth Analysis of Privacy Attitudes and Privacy Behaviors”. In: *European Journal of Social Psychology* 45.3 (2015), pp. 285–297.

[58] John R Douceur. “The Sybil Attack”. In: *International workshop on peer-to-peer systems*. Springer. 2002, pp. 251–260.

[59] Benjamin Dowling and Kenneth G Paterson. “A Cryptographic Analysis of the WireGuard Protocol”. In: *International Conference on Applied Cryptography and Network Security*. Springer. 2018, pp. 3–21.

[60] Cynthia Dwork. “Differential privacy”. In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2006, pp. 1–12.

[61] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, March 4-7, 2006, Proceedings*. New York, USA: Springer, 2006, pp. 265–284. DOI: 10.1007/11681878_14. URL: https://doi.org/10.1007/11681878_14.

[62] Nico Ebert, Kurt Alexander Ackermann, and Peter Heinrich. “Does Context in Privacy Communication Really Matter? — A Survey on Consumer Concerns and Preferences”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: ACM, 2020, 1–11. ISBN: 9781450367080. DOI: 10.1145/3313831.3376575. URL: https://doi.org/10.1145/3313831.3376575.

[63] Khaled El Emam, Saeed Samet, Luk Arbuckle, Robyn Tamblyn, Craig Earle, and Murat Kantarcioglu. “A Secure Distributed Logistic Regression Protocol for the Detection of Rare Adverse Drug Events”. In: *Journal of the American Medical Informatics Association* 20.3 (2013), pp. 453–461.

[64] Carl M. Ellison. “Ceremony Design and Analysis”. In: *IACR Cryptology ePrint Archive* 2007 (2007), p. 399. URL: http://eprint.iacr.org/2007/399.

[65] José Estrada-Jiménez, Javier Parra-Arnau, Ana Rodríguez-Hoyos, and Jordi Forné. “Online Advertising: Analysis of Privacy Threats and Protection Approaches”. In: *Computer Communications* 100 (2017), pp. 32–51.

[66] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. “Large-Scale Readability Analysis of Privacy Policies”. In: *Proceedings of the International Conference on Web Intelligence*. 2017, pp. 18–25.

[67] Federal Trade Commission. *Android Flashlight App Developer Settles FTC Charges It Deceived Consumers*. https://goo.gl/Zf18jI. Accessed 2019-08-09. 2013. (Visited on 12/05/2013).

[68] Paul Feldman. “A practical scheme for non-interactive verifiable secret sharing”. In: *Annual Symposium on Foundations of Computer Science (Proceedings)* (Nov. 1987), pp. 427–438. DOI: 10.1109/SFCS.1987.4.

[69] Yuanyuan Feng, Yaxing Yao, and Norman Sadeh. “A Design Space for Privacy Choices: Towards Meaningful Privacy Control in the Internet of Things”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–16.

[70] Diogo AB Fernandes, Liliana FB Soares, João V Gomes, Mário M Freire, and Pedro RM Inácio. “Security Issues in Cloud Environments: A Survey”. In: *International Journal of Information Security* 13.2 (2014), pp. 113–170.

[71] Casey Fiesler and Blake Hallinan. ““We Are the Product” Public Reactions to Online Data Sharing and Privacy Controversies in the Media”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–13.

[72] Daniel Franzen, Saskia Nuñez von Voigt, Peter Sörries, Florian Tschorsch, and Claudia Müller-Birn. “Am I Private and If So, how Many? Communicating Privacy Guarantees of Differential Privacy with Risk Communication Formats”. In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. New York, New York: ACM, 2022, pp. 1125–1139.

[73] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model inversion attacks that exploit confidence information and basic countermeasures”. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM. ACM, 2015, pp. 1322–1333.

[74] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, October 12-16, 2015*. Denver, CO, USA: ACM, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677. URL: https://doi.org/10.1145/2810103.2813677.

[75] Michael J. Freedman, Kobbi Nissim, and Benny Pinkas. “Efficient Private Matching and Set Intersection”. In: *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, May 2-6, 2004, Proceedings*. Interlaken, Switzerland: IACR, 2004, pp. 1–19. DOI: 10.1007/978-3-540-24676-3_1. URL: https://doi.org/10.1007/978-3-540-24676-3_1.

[76] Freedom of the Press Foundation. *Sunder is a user-friendly graphical interface for Shamir's Secret Sharing*. https://github.com/freedomofpress/sunder. Accessed 2019-05-28. 2018. (Visited on 10/13/2018).

[77] Freedom of the Press Foundation. *Welcome to Sunder*. https://sunder.readthedocs.io/en/latest/. Accessed 2019-05-28. 2018. (Visited on 10/13/2018).

[78] Tilman Frosch, Christian Mainka, Christoph Bader, Florian Bergsma, Jörg Schwenk, and Thorsten Holz. "How secure is TextSecure?" In: *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. IEEE. 2016, pp. 457–472.

[79] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. "The Limitations of Federated Learning in Sybil Settings". In: *the 23rd International Symposium on Research in Attacks, Intrusions and Defenses*. 2020.

[80] Ryan Gallagher and Glenn Greenwald. "How the NSA Plans to Infect 'Millions' of Computers with Malware". In: *The Intercept* (Mar. 12, 2014). URL: https://theintercept.com/2014/03/12/nsa-plans-infect-millions-computers-malware/.

[81] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, October 15-19, 2018*. Toronto, Canada: ACM, 2018, pp. 619–633. DOI: 10.1145/3243734.3243834. URL: https://doi.org/10.1145/3243734.3243834.

[82] Chaim Gartenberg. *Google Buys Fitbit for $2.1 Billion*. The Verge. https://www.theverge.com/2019/11/1/20943318/google-fitbit-acquisition-fitness-tracker-announcement.

[83] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. "Inverting gradients-how easy is it to break privacy in federated learning?" In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 16937–16947.

[84] Rosario Gennaro, Stanislaw Jarecki, Hugo Krawczyk, and Tal Rabin. "Robust Threshold DSS Signatures". In: *EUROCRYPT*. 1996, pp. 354–371.

[85] Robin C Geyer, Tassilo Klein, and Moin Nabi. "Differentially Private Federated Learning: A Client Level Perspective". In: *arXiv preprint arXiv:1712.07557* (2017).

[86] Oded Goldreich, Silvio Micali, and Avi Wigderson. "How to Play any Mental Game or A Completeness Theorem for Protocols with Honest Majority". In: *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987*. New York, USA: ACM, 1987, pp. 218–229. DOI: 10.1145/28395.28420. URL: https://doi.org/10.1145/28395.28420.

[87] Adam Groce, Peter Rindal, and Mike Rosulek. "Cheaper Private Set Intersection via Differentially Private Leakage". In: *Proceedings on Privacy Enhancing Technologies* 2019.3 (2019), pp. 6–25.

[88] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. ""It's a Scavenger Hunt": Usability of Websites' Opt-Out and Data Deletion Choices". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12.

[89] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. "Toggles, Dollar Signs, and Triangles: How to (in) effectively Convey Privacy Choices with Icons and Link Texts". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–25.

[90] Eszter Hargittai and Elissa M Redmiles. "Will Americans be Willing to Install COVID-19 Tracking Apps?" In: *Scientific American* (2020), Epub–ahead.

[91] Jamie Hayes and Olga Ohrimenko. "Contamination Attacks and Mitigation in Multi-Party Machine Learning". In: *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 6604–6615.

[92] Amir Herzberg, Stanisław Jarecki, Hugo Krawczyk, and Moti Yung. "Proactive Secret Sharing Or: How to Cope With Perpetual Leakage". In: *Advances in Cryptology — CRYPT0' 95*. Ed. by Don Coppersmith. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 339–352.

[93] Anis Heydari. *The Canadian Tech Company that Changed its Mind about Using Your Tax Return to Sell Stuff*. CBC. https://www.cbc.ca/radio/costofliving/indigenous-nations-and-the-economy-plus-why-it-s-so-hard-to-fly-for-cheap-in-canada-1.5469919/the-canadian-tech-company-that-changed-its-mind-about-using-your-tax-return-to-sell-stuff-1.5471400. 2020.

[94] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. "Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning". In: *the 2017 ACM SIGSAC Conference on Computer and Communications Security*. Dallas, TX, USA: ACM, 2017, pp. 603–618.

[95] Briland Hitaj, Giuseppe Ateniese, and Fernando Pérez-Cruz. "Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, October 30 - November 03, 2017*. Dallas, TX, USA: ACM, 2017, pp. 603–

618. DOI: 10.1145/3133956.3134012. URL: https://doi.org/10.1145/3133956.3134012.

[96] Karen Holtzblatt and Hugh Beyer. *Contextual Design: Defining Customer-Centered Systems*. Oxford, United Kingdom: Elsevier, 1997.

[97] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. "Membership Inference Attacks on Machine Learning: A Survey". In: *ACM Comput. Surv.* (2022). ISSN: 0360-0300. DOI: 10.1145/3523273. URL: https://doi.org/10.1145/3523273.

[98] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. "Source Inference Attacks in Federated Learning". In: *2021 IEEE International Conference on Data Mining (ICDM)*. 2021, pp. 1102–1107. DOI: 10.1109/ICDM51629.2021.00129.

[99] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. "Practical Blind Membership Inference Attack via Differential Comparisons". In: *Network and Distributed System Security Symposium (NDSS)* (2021).

[100] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. "Investigating membership inference attacks under data dependencies". In: *2023 2023 IEEE 36th Computer Security Foundations Symposium (CSF)(CSF)*. IEEE Computer Society, 2023, pp. 194–209.

[101] Paul Irolla and Grégory Châtel. "Demystifying the membership inference attack". In: *2019 12th CMI Conference on Cybersecurity and Privacy (CMI)*. IEEE. 2019, pp. 1–7.

[102] Markus Jakobsson. "The human factor in phishing". In: *Privacy & Security of Consumer Information* 7.1 (2007), pp. 1–19.

[103] Javelin Strategy & Research. *2017 State of Authentication Report*. https://fido-alliance-.org/wp-content/uploads/The-State-of-Authentication-Report.pdf. 2017.

[104] Bargav Jayaraman and David Evans. "Evaluating Differentially Private Machine Learning in Practice". In: *the 28th USENIX Security Symposium*. Santa Clara, CA, 2019, pp. 1895–1912.

[105] Xue Jiang, Xuebing Zhou, and Jens Grossklags. "Comprehensive Analysis of Privacy Leakage in Vertical Federated Learning During Prediction." In: *Proc. Priv. Enhancing Technol.* 2022.2 (2022), pp. 263–281.

[106] Arthur Jochems, Timo M Deist, Issam El Naqa, Marc Kessler, Chuck Mayo, Jackson Reeves, Shruti Jolly, Martha Matuszak, Randall Ten Haken, Johan van Soest, et al. "Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries". In: *International Journal of Radiation Oncology Biology Physics* 99.2 (2017), pp. 344–352.

[107] Arthur Jochems, Timo M Deist, Johan Van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. "Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept". In: *Radiotherapy and Oncology* 121.3 (2016), pp. 459–467.

[108] Bailey Kacsmar, Basit Khurram, Nils Lukas, Alexander Norton, Masoumeh Shafieinejad, Zhiwei Shang, Yaser Baseri, Maryam Sepehri, Simon Oya, and Florian Kerschbaum. "Differentially Private Two-Party Set Operations". In: *2020 IEEE European Symposium on Security and Privacy*. IEEE. 2020, pp. 390–404.

[109] Bailey Kacsmar, Chelsea Komlo, Florian Kerschbaum, and Ian Goldberg. "Mind the Gap: Ceremonies for Applied Secret Sharing." In: *Proc. Priv. Enhancing Technol.* 2020.2 (2020), pp. 397–415.

[110] Bailey Kacsmar, Kyle Tilbury, Miti Mazmudar, and Florian Kerschbaum. "Caring about Sharing: User Perceptions of Multiparty Data Sharing". In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 899–916. ISBN: 978-1-939133-31-1. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/kacsmar.

[111] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. ""My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security". In: *Eleventh Symposium On Usable Privacy and Security 2015)*. 2015, pp. 39–52.

[112] Farzaneh Karegar, Ala Sarah Alaqra, and Simone Fischer-Hübner. "Exploring User-Suitable Metaphors for Differentially Private Data Analyses". In: *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. Boston, MA: USENIX Association, Aug. 2022, pp. 175–193. ISBN: 978-1-939133-30-4. URL: https://www.usenix.org/conference/soups2022/presentation/karegar.

[113] Sowmya Karunakaran, Kurt Thomas, Elie Bursztein, and Oxana Comanescu. "Data Breaches: User Comprehension, Expectations, and Concerns with Handling Exposed Data". In: *Fourteenth Symposium on Usable Privacy and Security 2018)*. 2018, pp. 217–234.

[114] Jiro Kawakita. "The Original KJ Method". In: *Tokyo: Kawakita Research Institute* 5 (1991).

[115] Yigitcan Kaya and Tudor Dumitras. "When Does Data Augmentation Help With Membership Inference Attacks?" In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5345–5355.

[116] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. "A" nutrition label" for privacy". In: *Proceedings of the 5th Symposium on Usable Privacy and Security*. 2009, pp. 1–12.

[117] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. "Privacy as Part of the App Decision-Making Process". In: *Proceedings of the CHI conference on human factors in computing systems*. 2013, pp. 3393–3402.

[118] Lea Kissner and Dawn Song. "Privacy-Preserving Set Operations". In: *Annual International Cryptology Conference*. Berlin Heidelberg: Springer, 2005, pp. 241–257.

[119] John Koetsier. *Apple's Ad Network Gets 'Preferential Access To Users' Data' vs Facebook, Google, Others*. Forbes. https://www.forbes.com/sites/johnkoetsier/2020/08/07/apple-ad-network-gets-special-privileges-that-facebook\-google-wont-on-ios14/. 2021.

[120] Spyros Kokolakis. "Privacy Attitudes and Privacy Behaviour: A Review of Current Research on the Privacy Paradox Phenomenon". In: *Computers & security* 64 (2017), pp. 122–134.

[121] Ivar Krumpal. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review". In: *Quality & Quantity* 47.4 (2013), pp. 2025–2047.

[122] Patrick Kühtreiber, Viktoriya Pak, and Delphine Reinhardt. "Replication: The Effect of Differential Privacy Communication on German Users' Comprehension and Data Sharing Attitudes". In: *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. Boston, MA: USENIX, 2022, pp. 117–134.

[123] Patrick Kühtreiber and Delphine Reinhardt. "Usable Differential Privacy for the Internet-of-Things". In: *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. Kassel, Germany: IEEE, 2021, pp. 426–427.

[124] Ponnurangam Kumaraguru and Lorrie Faith Cranor. *Privacy indexes: a survey of Westin's studies*. Carnegie Mellon University, School of Computer Science, Institute for ..., 2005.

[125] Hongkyu Lee, Jeehyeong Kim, Seyoung Ahn, Rasheed Hussain, Sunghyun Cho, and Junggab Son. "Digestive neural networks: A novel defense strategy against inference attacks in federated learning". In: *computers & security* 109 (2021), p. 102378.

[126] Ed Leefeldt and Amy Danise Ed. *The Witness Against You: Your Car*. Forbes. https://www.forbes.com/advisor/car-insurance/telematics-data-privacy/. 2021.

[127] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. "Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2012, pp. 589–598.

[128] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. "Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy Through Crowdsourcing". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM. New York, New York, USA: ACM, 2012, pp. 501–510.

[129] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I Hong. "Modeling Users' Mobile App Privacy Preferences: Restoring Usability in a Sea of Permission Settings". In: *Tenth Symposium On Usable Privacy and Security 2014)*. 2014, pp. 199–212.

[130] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. "The Privacy Policy Landscape After the GDPR". In: *arXiv preprint arXiv:1809.08396* (2018).

[131] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. "ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models". In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/liu-yugeng.

[132] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. "A pragmatic approach to membership inferences on machine learning models". In: *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2020, pp. 521–534.

[133] Isis Agora Lovecruft and Henry De Valence. *https://doc.dalek.rs/curve25519_dalek/*. 2018. URL: https://doc.dalek.rs/curve25519_dalek/ (visited on 10/13/2018).

[134] Hermann C Lythgoe and Daniel S Dunn. "Food and Drug Legislation in Massachusetts". In: *Food Drug Cosm. LQ* 3 (1948), p. 565.

[135] Maureen Mahoney. *California Consumer Privacy Act: Are Consumers' Digital Rights Protected*. Tech. rep. https://advocacy.consumerreports.org/press_release/consumer-reports-study-finds-significant-obstacles-to-exercising-california-privacy-rights/. Technical Report. Consumer Reports., 2020.

[136] Francesco M. Malvestuto, Mauro Mezzini, and Marina Moscarini. "Auditing Sum-Queries to Make a Statistical Database Secure". In: *ACM Trans. Inf. Syst. Secur.* 9.1 (2006), 31–60. ISSN: 1094-9224. DOI: 10.1145/1127345.1127347. URL: https://doi.org/10.1145/1127345.1127347.

[137] Taciane Martimiano, Jean Everson Martina, M Maina Olembo, and Marcelo Carlomagno Carlos. "Modelling user devices in security ceremonies". In: *2014 Workshop on Socio-Technical Aspects in Security and Trust*. IEEE. 2014, pp. 16–23.

[138] Jean Everson Martina, Túlio Cícero Salavaro de Souza, and Ricardo Felipe Custodio. "Ceremonies Formal Analysis in PKI's Context". In: *2009 International Conference on Computational Science and Engineering*. Vol. 3. IEEE. 2009, pp. 392–398.

[139] Peter Mayer, Yixin Zou, Florian Schaub, and Adam J Aviv. ""Now I'm a Bit Angry:" Individuals' Awareness, Perception, and Responses to Data Breaches that Affected Them". In: *The Thirtieth USENIX Security Symposium 2021*. 2021.

[140] Aleecia M McDonald and Lorrie Faith Cranor. "Americans' Attitudes About Internet Behavioral Advertising Practices". In: *Proceedings of the Ninth Annual ACM Workshop on Privacy in the Electronic Society*. 2010, pp. 63–72.

[141] Aleecia M McDonald and Lorrie Faith Cranor. "The Cost of Reading Privacy Policies". In: *ISJLP* 4 (2008), p. 543.

[142] Chris McGreal. "Martin Luther King friend and photographer was FBI informant". In: *The Guardian* (Sept. 14, 2010). URL: https://www.theguardian.com/world/2010/sep/14/photographer-ernest-withers-fbi-informer (visited on 11/18/2018).

[143] Susan E. McGregor, Elizabeth Anne Watkins, Mahdi Nasrullah Al-Ameen, Kelly Caine, and Franziska Roesner. "When the Weakest Link is Strong: Secure Collaboration in the Case of the Panama Papers". In: *26th USENIX Security Symposium (USENIX Security 2017)*. Vancouver, BC: USENIX Association, 2017, pp. 505–522. ISBN: 978-1-931971-40-9. URL: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/mcgregor.

[144] James McLeod. *Double-Double Tracking: How Tim Hortons Knows Where You Sleep, Work and Vacation*. Financial Post. https://financialpost.com/technology/tim-hortons-app-tracking-customers-intimate-data. 2020.

[145] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *the 20th International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale, FL, USA: PMLR, 2017, pp. 1273–1282.

[146] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017*. Fort Lauderdale, USA: ACM, 2017, pp. 1273–1282. URL: http://proceedings.mlr.press/v54/mcmahan17a.html.

[147] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. "Learning Differentially Private Recurrent Language Models". In: *International Conference on Learning Representations*. 2018.

[148] Catherine Meadows. "A more efficient cryptographic matchmaking protocol for use in the absence of a continuously available third party". In: *1986 IEEE Symposium on Security and Privacy*. IEEE. 1986, pp. 134–134.

[149] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. "Exploiting unintended feature leakage in collaborative learning". In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, pp. 691–706.

[150] Ute Mielow-Weidmann and Paul Weidmann. "Das Bundesdatenschutzgesetz". In: *Wirtschafts-, Rechts- und Sozialkunde für Sekretärinnen*. Wiesbaden: Gabler Verlag, 1996, pp. 287–289. ISBN: 978-3-663-05816-8. DOI: 10.1007/978-3-663-05816-8_19. URL: https://doi.org/10.1007/978-3-663-05816-8_19.

[151] Matthew B Miles, A Michael Huberman, and Johnny Saldaña. *Qualitative Data Analysis: A Methods Sourcebook*. Sage publications, 2018.

[152] Ambar Murillo, Andreas Kramm, Sebastian Schnorf, and Alexander De Luca. ""If I Press Delete, It's Gone"-User Understanding of Online Data Deletion and Expiration". In: *Fourteenth Symposium on Usable Privacy and Security 2018)*. 2018, pp. 329–339.

[153] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. "Privacy Expectations and Preferences in an IoT World". In: *Thirteenth Symposium on Usable Privacy and Security 2017*. 2017, pp. 399–412.

[154] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. "Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases". In: *Proceedings on Privacy Enhancing Technologies* 2 (2022), pp. 601–618.

[155] Arvind Narayanan, Joanna Huey, and Edward W Felten. "A Precautionary Approach to Big Data Privacy". In: *Data Protection on the Move*. Springer, 2016, pp. 357–385.

[156] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks". In: *arXiv preprint arXiv:1812.00910* (2018).

[157] Ventzislav Nikov and Svetla Nikova. "On Proactive Secret Sharing Schemes". In: *International Workshop on Selected Areas in Cryptography*. Springer. 2004, pp. 308–325.

[158] Helen Nissenbaum. "Contextual integrity up and down the data food chain". In: *Theoretical Inquiries in Law* 20.1 (2019), pp. 221–256.

[159] Thomas B Norton. "The Non-Contractual Nature of Privacy Policies and a New Critique of the Notice and Choice Privacy Protection Model". In: *Fordham Intell. Prop. Media & Ent. LJ* 27 (2016), p. 181.

[160] Maggie Oates, Yama Ahmadullah, Abigail Marsh, Chelse Swoopes, Shikun Zhang, Rebecca Balebako, and Lorrie Faith Cranor. "Turtles, Locks, and Bathrooms: Understanding Mental Models of Privacy Through Illustration". In: *Proceedings on Privacy Enhancing Technologies* 2018.4 (2018), pp. 5–32.

[161] Jonathan A Obar and Anne Oeldorf-Hirsch. "The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services". In: *Information, Communication & Society* (2018), pp. 1–20.

[162] Sean O'Connor, Ryan Nurwono, and Eleanor Birrell. "(Un) clear and (In) conspicuous: The Right to Opt-Out of Sale Under CCPA". In: *arXiv preprint arXiv:2009.07884* (2020).

[163] Sean O'Connor, Ryan Nurwono, Aden Siebel, and Eleanor Birrell. "(Un) clear and (In) conspicuous: The right to opt-out of sale under CCPA". In: *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*. New York, New York: ACM, 2021, pp. 59–72.

[164] Office of the Privacy Commissioner of Canada. *PIPEDA in brief*. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/. Accessed 2019-06-18. 2019. (Visited on 05/2019).

[165] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The Measurement of Meaning*. University of Illinois press, 1957.

[166] Rafail Ostrovsky and Moti Yung. "How to Withstand Mobile Virus Attacks (Extended Abstract)". In: *Proceedings of the Tenth Annual ACM Symposium on Principles of Distributed Computing*. PODC '91. Montreal, Quebec, Canada: ACM, 1991, pp. 51–59. ISBN: 0-89791-439-2. DOI: 10.1145/112600.112605. URL: http://doi.acm.org/10.1145/112600.112605.

[167] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data". In: *the International Conference on Learning Representations*. Toulon, France, 2017.

[168] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. "SoK: Security and Privacy in Machine Learning". In: *the 2018 IEEE European Symposium on Security and Privacy*. IEEE. London, UK, 2018, pp. 399–414.

[169] Jaehong Park and Ravi Sandhu. "A Position Paper: A Usage Control (UCON) Model for Social Networks Privacy". In: (2000).

[170] Jaehong Park and Ravi Sandhu. "The UCON$_{ABC}$ Usage Control Model". In: *ACM Trans. Inf. Syst. Secur.* 7.1 (Feb. 2004), 128–174. ISSN: 1094-9224. DOI: 10.1145/984334.984339. URL: https://doi.org/10.1145/984334.984339.

[171] Torben P. Pedersen. "Non-Interactive and Information-Theoretic Secure Verifiable Secret Sharing". In: *Proceedings of the 11th Annual International Cryptology Conference on Advances in Cryptology*. CRYPTO '91. London, UK, UK: Springer-Verlag, 1992, pp. 129–140. ISBN: 3-540-55188-3. URL: http://dl.acm.org/citation.cfm?id=646756.705507.

[172] Trevor Perrin and Moxie Marlinspike. *The Double Ratchet Algorithm*. https://signal.org/docs/specifications/doubleratchet/. 2016. (Visited on 08/14/2019).

[173] Sandra Petronio. *Boundaries of privacy: Dialectics of disclosure*. Suny Press, 2002.

[174] Benny Pinkas, Thomas Schneider, Gil Segev, and Michael Zohner. "Phasing: Private Set Intersection Using Permutation-based Hashing". In: *24th USENIX Security Symposium (USENIX Security 15)*. Washington, D.C.: USENIX Association, Aug. 2015, pp. 515–530. ISBN: 978-1-939133-11-3. URL: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/pinkas.

[175] Benny Pinkas, Thomas Schneider, Christian Weinert, and Udi Wieder. "Efficient Circuit-Based PSI via Cuckoo Hashing". In: *Advances in Cryptology – EUROCRYPT 2018*. Ed. by Jesper Buus Nielsen and Vincent Rijmen. Cham: Springer International Publishing, 2018, pp. 125–157.

[176] Benny Pinkas, Thomas Schneider, and Michael Zohner. "Scalable Private Set Intersection Based on OT Extension". In: *ACM Trans. Priv. Secur.* 21.2 (2018). ISSN: 2471-2566. DOI: 10.1145/3154794. URL: https://doi.org/10.1145/3154794.

[177] Stanley Presser, Mick P Couper, Judith T Lessler, Elizabeth Martin, Jean Martin, Jennifer M Rothgeb, and Eleanor Singer. "Methods for Testing and Evaluating Survey Questions". In: *Public opinion quarterly* 68.1 (2004), pp. 109–130.

[178] Lucy Qin, Andrei Lapets, Frederick Jansen, Peter Flockhart, Kinan Dak Albab, Ira Globus-Harris, Shannon Roberts, and Mayank Varia. "From usability to secure computing and back again". In: *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*. 2019, pp. 191–210.

[179] Lucy Qin, Andrei Lapets, Frederick Jansen, Peter Flockhart, Kinan Dak Albab, Ira Globus-Harris, Shannon Roberts, and Mayank Varia. "From Usability to Secure Computing and Back Again". In: *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 191–210. ISBN: 978-1-939133-05-2. URL: https://www.usenix.org/conference/soups2019/presentation/qin.

[180] Emilee Rader. "Awareness of Behavioral Tracking and Information Privacy Concern in Facebook and Google". In: *Tenth Symposium On Usable Privacy and Security 2014*. 2014, pp. 51–67.

[181] Kenneth Radke, Colin Boyd, Juan Gonzalez Nieto, and Margot Brereton. "Ceremony analysis: Strengths and weaknesses". In: *IFIP International Information Security Conference*. Springer. 2011, pp. 104–115.

[182] Anjana Rajan, Lucy Qin, David W Archer, Dan Boneh, Tancrede Lepoint, and Mayank Varia. "Callisto: A cryptographic approach to detecting serial perpetrators of sexual misconduct". In: *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM. 2018, p. 49.

[183] Joel Reardon. *Secure Data Deletion*. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-28778-2. DOI: 10.1007/978-3-319-28778-2.

[184] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. "Hogwild: A lock-free approach to parallelizing stochastic gradient descent". In: *Advances in neural information processing systems*. 2011, pp. 693–701.

[185] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. "How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples". In: *2019 IEEE Symposium on Security and Privacy (SP)*. 2019, pp. 1326–1343. DOI: 10.1109/SP.2019.00014.

[186] Elissa M Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L Mazurek. "Asking for a Friend: Evaluating Response Biases in Security User Studies". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM. New York, New York, USA: ACM, 2018, pp. 1238–1255.

[187] Joel R Reidenberg, N Cameron Russell, Alexander J Callen, Sophia Qasir, and Thomas B Norton. "Privacy Harms and the Effectiveness of the Notice and Choice Framework". In: *ISJLP* 11 (2015), p. 485.

[188] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. "Why doesn't Jane protect her privacy?" In: *International Symposium on Privacy Enhancing Technologies Symposium*. Springer. 2014, pp. 244–262.

[189] Eric Rescorla. *The Transport Layer Security (TLS) Protocol Version 1.3*. https://tools.ietf.org/html/rfc8446. ID. 2018.

[190] John A Rothchild. "Against Notice and Choice: The Manifest Failure of the Proceduralist Paradigm to Protect Privacy Online (Or Anywhere Else)". In: *Clev. St. L. Rev.* 66 (2017), p. 559.

[191] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. "White-box vs black-box: Bayes optimal strategies for membership inference". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5558–5567.

[192] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. "Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning". In: *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020, pp. 1291–1308. ISBN: 978-1-939133-17-5. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/salem.

[193] Jerome H. Saltzer and Michael D. Schroeder. "The protection of information in computer systems". In: *Proceedings of the IEEE* 63.9 (1975), pp. 1278–1308. ISSN: 0018-9219. DOI: 10.1109/PROC.1975.9939.

[194] Florian Schaub, Rebecca Balebako, and Lorrie Faith Cranor. "Designing effective privacy notices and controls". In: *IEEE Internet Computing* 21.3 (2017), pp. 70–77.

[195] Christophe Olivier Schneble, Bernice Simone Elger, and David Martin Shaw. "Google's Project Nightingale Highlights the Necessity of Data Science Ethics Review". In: *EMBO molecular medicine* 12.3 (2020), e12053.

[196] Bruce Schneier. *Cell Phone Opsec*. https://www.schneier.com/blog/archives/2015/04/cell_phone_opse.html. Accessed 2019-05-24. 2019.

[197] Bruce Schneier. *The Operating System That Can Protect You Even if You Get Hacked*. https://freedom.press/news/the-operating-system-that-can-protect-you-even-if-you-get-hacked/. Accessed 2019-05-14. Apr. 10, 2014.

[198] Jonathan Schulz, Duman Bahrami-Rad, Jonathan Beauchamp, and Joseph Henrich. "The Origins of WEIRD Psychology". In: *Available at SSRN 3201031* (2018).

[199] Raymond Scupin. "The KJ Method: A Technique for Analyzing Data Derived from Japanese Ethnology". In: *Human organization* 56.2 (1997), pp. 233–237.

[200] Adi Shamir. "How to share a secret". In: *Communications of the ACM* 22 (1979), pp. 612–613.

[201] Tariq Shaukat. *Our partnership with Ascension*. Google Cloud. https://cloud.google.com/blog/topics/inside-google-cloud/our-partnership-with-ascension. Google Cloud, 2019.

[202] Virat Shejwalkar, Huseyin A Inan, Amir Houmansadr, and Robert Sim. "Membership Inference Attacks Against NLP Classification Models". In: *NeurIPS 2021 Workshop Privacy in Machine Learning*. 2021.

[203] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. "Multi-institutional Deep Learning Modeling without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation". In: *International MICCAI Brainlesion Workshop*. Springer. Cham, 2018, pp. 92–104.

[204] Irina Shklovski, Scott D Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. "Leakiness and Creepiness in App Space: Perceptions of Privacy and Mobile App Use". In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM. New York, New York, USA: ACM, 2014, pp. 2347–2356.

[205] Reza Shokri and Vitaly Shmatikov. "Privacy-Preserving Deep Learning". In: *the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM. Denver, CO, USA, 2015, pp. 1310–1321.

[206] Reza Shokri, Martin Strobel, and Yair Zick. "On the privacy risks of model explanations". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 231–241.

[207] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks Against Machine Learning Models". In: *the 2017 IEEE Symposium on Security and Privacy*. IEEE. San Jose, CA, USA, 2017, pp. 3–18.

[208] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks Against Machine Learning Models". In: *2017 Symposium on Security and Privacy, SP 2017, May 22-26, 2017*. San Jose, USA: IEEE, 2017, pp. 3–18. DOI: 10.1109/SP.2017.41. URL: https://doi.org/10.1109/SP.2017.41.

[209] Santiago Silva, Boris A Gutman, Eduardo Romero, Paul M Thompson, Andre Altmann, and Marco Lorenzi. "Federated Learning in Distributed Medical Databases: Meta-analysis of Large-scale Subcortical Brain Data". In: *the 2019 IEEE 16th International Symposium on Biomedical Imaging*. IEEE. Venice, Italy, 2019, pp. 270–274.

[210] Jesper Simonsen and Toni Robertson. *Routledge International Handbook of Participatory Design*. Vol. 711. New York: Routledge New York, 2013.

[211] Mary Anne Smart, Dhruv Sood, and Kristin Vaccaro. "Understanding Risks of Privacy Theater with Differential Privacy". In: *CSCW 2022*. New York, New York: ACM, 2022.

[212] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. "Federated Multi-Task Learning". In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4424–4434.

[213] Congzheng Song and Vitaly Shmatikov. "Auditing data provenance in text-generation models". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 196–206.

[214] Liwei Song and Prateek Mittal. "Systematic evaluation of privacy risks of machine learning models". In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 2615–2632.

[215] Liwei Song, Reza Shokri, and Prateek Mittal. "Membership Inference Attacks Against Adversarially Robust Deep Learning Models". In: *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE Computer Society. 2019, pp. 50–56.

[216] Liwei Song, Reza Shokri, and Prateek Mittal. "Privacy risks of securing machine learning models against adversarial examples". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019, pp. 241–257.

[217] Spin Research. *https://github.com/SpinResearch/RustySecrets*. 2018. URL: https://github.com/SpinResearch/RustySecrets (visited on 12/01/2018).

[218] Frank Stajano. "Pico: No More Passwords!" In: *Proceedings of the 19th International Conference on Security Protocols*. SP'11. Cambridge, UK: Springer-Verlag, 2011, pp. 49–81. ISBN: 978-3-642-25866-4. DOI: 10.1007/978-3-642-25867-1_6. URL: http://dx.doi.org/10.1007/978-3-642-25867-1_6.

[219] State of California Department of Justice. *California Consumer Privacy Act*. https://oag.ca.gov/privacy/ccpa. Accessed 2022-09-04. 2018. (Visited on 2018).

[220] Peter Story, Daniel Smullen, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. "Awareness, Adoption, and Misconceptions of Web Privacy Tools". In: *Proceedings on Privacy Enhancing Technologies* 3 (2021), pp. 308–333.

[221] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. ""I Don't Own the Data": End User Perceptions of Smart Home Device Data Practices and Risks". In: *Fifteenth Symposium on Usable Privacy and Security 2019*. 2019.

[222] TensorFlow. *TensorFlow Federated: Machine Learning on Decentralized Data*. https://www.tensorflow.org/federated. Accessed 2022. 2022.

[223] Igor V Tetko, David J Livingstone, and Alexander I Luik. "Neural network studies. 1. Comparison of overfitting and overtraining". In: *Journal of chemical information and computer sciences* 35.5 (1995), pp. 826–833.

[224] StartUp HERE Toronto. *Waterloo-Based Bonfire Acquired for $140 Million CAD in 'Govtech' Merger*. StartUp HERE Toronto. https://startupheretoronto.com/partners/betakit/waterloo-based-bonfire-acquired-for-140-million-cad\-in-govtech-merger/. 2018.

[225] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. "A Hybrid Approach to Privacy-Preserving Federated Learning". In: *the 12th ACM Workshop on Artificial Intelligence and Security*. London, England: ACM, 2019, pp. 1–11.

[226] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. "Demystifying membership inference attacks in machine learning as a service". In: *IEEE Transactions on Services Computing* (2019).

[227] Janice Y Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. "The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study". In: *Information Systems Research* 22.2 (2011), pp. 254–268.

[228] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. "Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising". In: *Eighth Symposium on Usable Privacy and Security 2012*. 2012, pp. 1–15.

[229] U.S. Federal Trade Commission. *Children's Online Privacy Protection Rule*. https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa. Accessed 2022-09-04. 2022. (Visited on 2022).

[230] Paul Voigt and Axel Von dem Bussche. "The EU General Data Protection Regulation (GDPR)". In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017).

[231] Alan F Westin. "Privacy and Freedom". In: (1967).

[232] Alma Whitten and J Doug Tygar. "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0." In: *USENIX Security Symposium*. Vol. 348. 1999.

[233] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. "The Feasibility of Dynamically Granted Permissions: Aligning Mobile Privacy with User Preferences". In: *2017 IEEE Symposium on Security and Privacy*. IEEE. 2017, pp. 1077–1093.

[234] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Laura Brandimarte, and Alessandro Acquisti. "Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened as a result? the westin categories, behavioral intentions, and consequences". In: *10th Symposium On Usable Privacy and Security 2014)*. 2014, pp. 1–18.

[235] Justin Wu and Daniel Zappala. "When is a tree really a truck? Exploring Mental Models of Encryption". In: *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. Baltimore, MD: USENIX, 2018, pp. 395–409.

[236] Yuncheng Wu, Shaofeng Cai, Xiaokui Xiao, Gang Chen, and Beng Chin Ooi. "Privacy Preserving Vertical Federated Learning for Tree-based Models". In: *Proceedings of the VLDB Endowment* 13.11 ().

[237] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. "Towards Effective Differential Privacy Communication for Users' Data Sharing Decision and Comprehension". In: *arXiv preprint arXiv:2003. 13922* (2020).

[238] Runhua Xu, Nathalie Baracaldo, Yi Zhou, Ali Anwar, and Heiko Ludwig. "HybridAlpha: An Efficient Approach for Privacy-Preserving Federated Learning". In: *the 12th ACM Workshop on Artificial Intelligence and Security*. London, England: ACM, 2019, pp. 13–23.

[239] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. "Applied Federated Learning: Improving Google Keyboard Query Suggestions". In: *arXiv preprint arXiv:1812.02903* (2018).

[240] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. "Neural network inversion in adversarial setting via background knowledge alignment". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS '19. London, United Kingdom: ACM, 2019, pp. 225–240. ISBN: 978-1-4503-6747-9. DOI: 10.1145/3319535.3354261. URL: http://doi.acm.org/10.1145/3319535.3354261.

[241] Andrew Chi-Chih Yao. "How to Generate and Exchange Secrets (Extended Abstract)". In: *27th Annual Symposium on Foundations of Computer Science, 27-29 October 1986*. Toronto, Canada: IEEE, 1986, pp. 162–167. DOI: 10.1109/SFCS.1986.25. URL: https://doi.org/10.1109/SFCS.1986.25.

[242] Yaxing Yao, Davide Lo Re, and Yang Wang. "Folk Models of Online Behavioral Advertising". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017, pp. 1957–1969.

[243] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. "Privacy risk in machine learning: Analyzing the connection to overfitting". In: *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE. 2018, pp. 268–282.

[244] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting". In: *31st Computer Security Foundations Symposium, CSF 2018, July 9-12, 2018*. Oxford, United Kingdom: IEEE, 2018, pp. 268–282. DOI: 10.1109/CSF.2018.00027. URL: https://doi.org/10.1109/CSF.2018.00027.

[245] Xun Yi, Russell Paulet, and Elisa Bertino. "Homomorphic encryption". In: *Homomorphic Encryption and Applications*. Springer, 2014, pp. 27–46.

[246] Xuefei Yin, Yanming Zhu, and Jiankun Hu. "A Comprehensive Survey of Privacy-Preserving Federated Learning: A Taxonomy, Review, and Future Directions". In: *ACM Comput. Surv.* 54.6 (2021). ISSN: 0360-0300. DOI: 10.1145/3460427. URL: https://doi.org/10.1145/3460427.

[247] Jingwen Zhang, Jiale Zhang, Junjun Chen, and Shui Yu. "Gan enhanced membership inference: A passive local attack in federated learning". In: *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE. 2020, pp. 1–6.

[248] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. "Parallelized stochastic gradient descent". In: *Advances in neural information processing systems*. NIPS, 2010, pp. 2595–2603.

Appendix A

Construction Details for Secret Sharing Instances

A.1 Shamir Secret Sharing Construction

Share Generation and Distribution

1. Let the secret space be $\mathcal{S} = GF(q)^{\ell}$, where q is a prime or a prime power, $q \geq n+1$, and $\ell \geq 1$. Let $s \in \mathcal{S}$ be the secret.
2. The dealer D selects $t-1$ values independently and uniformly at random from $\mathcal{S}$ as $r_1, \ldots, r_{t-1}$.
3. The dealer computes $f : GF(q) \rightarrow \mathcal{S}$ as $f(x) = r_{t-1}\, x^{t-1} + r_{t-2}\, x^{t-2} + \cdots + r_1\, x + s$.
4. The dealer generates shares $s_i = (a_i, f(a_i))$ for $1 \leq i \leq n$, where the a_i are arbitrary distinct non-zero elements of $GF(q)$.
5. The dealer distributes $s_i = (a_i, f(a_i))$ to participant P_i for $1 \leq i \leq n$.
6. Delete r_i's from the device.

Reconstruction

1. A coalition of t participants combines their shares $\langle s_i = (x_i, y_i)\rangle_{i=1}^{t}$ and performs polynomial interpolation to recover the secret $s = f(0) = \sum_{j=1}^{t} y_j \prod_{1 \leq h \leq t, h \neq j} \frac{x_h}{x_h - x_j}$.

A.2 A Ceremony for Sunder

In the following we define a ceremony based on the information contained within the documentation for Sunder [77]. We use our framework to define the ceremony included in our analysis of Sunder. Note that we have re-categorized steps from our framework, for example device to action, depending on Sunder's implementation of each step.

A.2.1 Base Mode Stages

Share Generation

1. Choice: The dealer chooses values for n and t.
2. Device: Generates a signature keypair.
3. Device: Generates n shares (of the secret $s \in GF(256)^{\ell}$) $s_i = (a_i, f(a_i))$ for $1 \leq i \leq n$, where the a_i are arbitrary distinct non-zero elements of $GF(256)$. The shares are signed with the signature key.
4. Action: Delete all copies of s and the signature key. (The device retains the public verification key.)

Share Distribution

1. Choice: Select n participants (possibly including the dealer).
2. Choice: Select a secure communication channel (in person, Signal, etc.).
3. Action: The dealer distributes $s_i = (a_i, f(a_i))$ along with the public verification key to participant P_i for $1 \leq i \leq n$.
4. Action: Delete each s_i from the dealer's device. Exception is if the dealer is a participant and keeps one share.
5. Choice: Each participant selects an appropriate storage mechanism for their share.
6. Action: Each participant stores their share in the selected storage mechanism.

Reconstruction

1. Choice: Select a communication channel to bring t or more shares together.
2. Not Defined: P_r and the contacted participants authenticate one another.
3. Choice: Contacted participants elect whether to proceed and participate in a reconstruction.
4. Action: If proceeding, a contacted participant sends their share to P_r.
5. Device: Checks the signature on the received shares.
6. Device: Combines the t or more valid shares using polynomial interpolation to recover the secret $s = f(0)$.

A.2.2 Extended Mode Stages

Secret Preparation

1. Action: Generate a secret key to be used as s.
2. Action: Encrypt $\mathcal{F}$ using s and an appropriate authenticated[1] encryption algorithm.
3. Action: Store the ciphertext.

Extended Reconstruction

1. Action: Acquire ciphertext from storage.
2. Action: Use recovered s to decrypt the ciphertext.

A.3 A Ceremony for Shatter Secrets

Shatter Secrets only has an Extended mode of operation, which we define below. Shatter Secrets uses devices with Near Field Communication (NFC) capability in order to secret share a key s that is used to encrypt a sensitive drive such as a laptop.

Secret Preparation

[1] Although we specify an authenticated encryption algorithm, this is not specified by Sunder.

1. Action: Generate a secret key to be used as s.
2. Action: Encrypt the sensitive drive using s.

Share Generation

1. Choice: The user chooses values for n and t.
2. Device: Generates a symmetric key.
3. Device: Generates n shares (of the secret $s \in GF(256)^{\ell}$) $s_i = (a_i, f(a_i))$ for $1 \leq i \leq n$, where the a_i are arbitrary distinct non-zero elements of $GF(256)$. The shares are encrypted using authenticated encryption using the symmetric key.
4. Action: Delete all copies of s. (The device retains the symmetric key.)

Share Distribution

1. Choice: Select n participants.
2. Choice: Select a secure communication channel (in person, Signal, etc.).
3. Action: The user distributes the encrypted share $E(s_i)$ to participant P_i for $1 \leq i \leq n$.
4. Action: Delete each s_i from the user's device.
5. Device: Each participant's device stores their share.

Reconstruction

1. Action: Meet participant with shares. NFC tap user's device to participant's device to transfer encrypted share. Repeat until t or more shares are retrieved.
2. Device: Decrypt each share with stored symmetric key, discarding unsuccessful decryptions.
3. Device: Combine the t or more valid shares using polynomial interpolation to recover the secret $s = f(0)$.
4. Device: Use recovered s to decrypt the sensitive drive.

Appendix B

Additional Details for Federated Machine Learning Experiments

B.1 Training federated learning

Global parameters. The total data used to train a target model is divided equally among clients without overlapping. Unless otherwise specified there is always one attack and $n-1$ honest participants, where n is the total number of participants. Models are trained using 50 global rounds and 5 local epochs. The server learning rate is set to 1.0, but for federated averaging this parameter does not really apply. The test batch sizes are 1028.

EMNIST specific parameters. The client learning rate is 0.1 with a default batch size of $b = 10$ unless otherwise specified. The model architecture follows McMahan et al. [145]. The model is a convolutional neural network (CNN) with two 5x5 convolution layers (the first with 32 channels, the second with 64, each followed with 2x2 max pooling), a fully connected layer with 512 units and ReLu activation, and a final softmax output layer (1,663,370 total parameters).

CIFAR specific parameters. The client learning rate is 0.05 and learning rate decay of 0.01 with a default batch size of $b = 64$ as per McMahan et al. [145]. The model architecture follows TensorFlow CNN https://www.tensorflow.org/tutorials/images/cnn which is the same as McMahan et al. [145]. The model is a convolutional neural network (CNN) with

Algorithm 2 Federated learning with model averaging [145]

```
/* Server executes: */
Initialize parameters θ_0
m ← size of subset of participating clients
for each iteration t do
    S_t ← random set of m clients
    for each client k ∈ S_t do
        θ_t^k ← ClientUpdate(θ_{t−1})
    θ_t ← Σ_k (n_k / n) θ_t^k
function ClientUpdate(θ)
    for each local epoch do
        for each batch B in client's data D do
            θ ← θ − η∇L(b; θ)
    return local updated θ
```

three 3x3 convolution layers (the first with 32 channels, the second and third with 64, each followed with 2x2 max pooling), a fully connected layer with 64 units and ReLu activation, and a final softmax output layer (122,570 total parameters).

Algorithms. Algorithms for performing federated learning with model averaging and selective batch stochastic gradient descent are included as Algorithm 2 and Algorithm 3.

Algorithm 3 Server synchronized selective batch SGD [149]

```
/* Server executes: */
Initialize parameters θ_0
for each Iteration t do
    for each client k do
        g_t^k ← ClientUpdate(θ_{t−1})
    θ_t ← θ_{t−1} − η Σ_k g_t^k
function ClientUpdate(θ)
    Select batch b from client's data D
    return local gradients ∇L(b; θ)
```

B.2 Expanded Results Data

In Table B.1, Table B.3 and Table B.3 we include an overview of our results; including configurations. We execute each configuration ten times, with the results consisting of the mean training and test accuracy results, the mean attack accuracy results, the standard deviations, and the confidence intervals.

EMNIST Results

$\|\mathcal{D}\|$	n	b	Training $\mu \pm$ CI	Testing $\mu \pm$ CI	Snapshot $\mu \pm$ CI	Distance $\mu \pm$ CI	Vanilla $\mu \pm$ CI	S-No Drop $\mu \pm$ CI
200	2	10	100±0	97±1	79±3	74±2	69±3	77±3
400	4	10	100±0	96±2	71±2	67±2	61±2	70±2
800	8	10	100±0	97±1	66±1	62±1	57±1	59±3
1600	16	10	100±0	97±1	59±2	53±4	53±1	65±1
2000	20	10	100±0	97±1	60±1	57±3	53±1	60±1
2400	24	10	99±0	97±1	58±1	56±3	52±1	59±1
2800	28	10	99±0	97±0	58±1	57±3	52±0	58±1
3200	32	10	99±0	97±0	58±1	55±3	52±0	58±1
400	2	10	100±0	96±1	74±2	67±2	63±2	74±3
800	2	10	100±0	98±1	68±2	63±1	59±1	68±1
1600	2	10	100±0	98±1	63±1	59±1	56±1	63±1
3200	2	10	100±0	98±0	58±1	55±0	53±0	58±1
6400	2	10	100±0	99±0	56±0	54±0	53±0	56±0
1600	8	10	100±0	97±1	62±1	59±2	54±1	62±1
3200	8	10	100±0	98±0	60±1	58±1	54±1	60±1
6400	8	10	100±0	98±0	57±0	55±0	52±0	57±0
3200	16	10	100±0	98±1	59±1	56±3	53±1	59±1
6400	16	10	100±0	98±0	57±1	56±1	52±0	57±1
200	2	40	100±0	95±4	70±3	76±3	66±2	49±3
300	3	40	100±0	95±3	66±3	67±4	61±2	53±3
400	4	40	100±0	96±2	64±3	64±2	59±2	53±2
800	8	40	99±0	96±1	62±1	54±3	55±1	56±2
1600	16	40	98±0	96±1	61±1	50±4	53±1	57±2
2400	24	40	97±0	96±1	59±1	49±2	51±0	57±1
200	2	80	100±0	95±3	62±5	63±3	63±3	39±2
300	3	80	99±0	96±2	62±4	61±3	61±3	39±2
400	4	80	99±1	94±2	62±3	59±1	59±1	42±5
800	8	80	97±1	94±1	61±2	54±1	54±1	41±5
1600	16	80	96±1	95±1	60±2	52±1	52±1	42±4
2400	24	80	95±1	94±1	57±2	51±1	51±1	41±2

Table B.1: The label S-No Drop designates a snapshot attack where no rounds are excluded from the majority vote. We present the mean accuracy and the 95% confidence intervals for the mean (CI).

	Snapshot		Distance		Vanilla	
n	**Reg.**	**Self**	**Reg.**	**Self**	**Reg.**	**Self**
2	79±3	80±4	74±2	75±3	69±3	69±5
4	71±2	72±3	67±2	67±2	61±2	61±2
8	66±1	68±3	62±1	64±2	57±1	58±2
16	60±3	59±6	59±2	57±7	53±1	55±3
20	60±1	59±2	57±3	56±5	53±1	54±1
24	58±2	58±3	56±3	57±6	52±1	53±2
28	58±1	58±2	57±3	57±6	52±0	53±2
32	58±1	58±2	55±3	53±5	52±0	52±1

Table B.2: Target models were trained on EMNIST for varying participating clients (n) and each client contributing 100 data points. Comparison of actual attack accuracy to the estimated accuracy that adversaries can compute. Self-attack accuracy is computed by executing the attack on their own training data.

CIFAR-10 Results

$\|\mathcal{D}\|$	**n**	**b**	Training $\boldsymbol{\mu \pm}$ **CI**	Testing $\boldsymbol{\mu \pm}$ **CI**	Snapshot $\boldsymbol{\mu \pm}$ **CI**	Distance $\boldsymbol{\mu \pm}$ **CI**	Vanilla $\boldsymbol{\mu \pm}$ **CI**
24000	2	64	64±2	62±2	66±0	48±2	51±0
24000	2	128	61±2	59±2	64±1	49±2	51±0
24000	4	64	62±2	60±2	65±1	50±1	51±0
24000	4	128	58±1	57±1	62±1	50±1	51±0
24000	8	64	59±1	58±1	62±1	50±0	51±0
24000	8	128	55±1	55±1	59±2	50±0	50±0
24000	16	64	55±1	54±1	59±1	50±0	50±0
24000	16	128	50±1	49±1	55±2	50±0	50±0
24000	20	64	53±1	53±1	55±2	50±0	50±0
24000	20	128	49±1	49±1	51±3	50±0	50±0
24000	24	64	52±1	51±1	52±3	50±0	50±0
24000	28	64	51±1	50±1	54±3	50±0	50±0
24000	28	128	45±1	46±1	51±3	50±0	50±0
10000	2	64	60±2	59±2	63±1	50±2	51±0
10000	2	128	57±1	57±1	61±1	50±1	51±0
10000	4	64	57±1	56±1	60±1	49±1	51±0
10000	4	128	53±1	53±1	58±2	50±0	50±0
10000	8	64	53±0	53±0	57±1	50±0	50±0
10000	8	128	49±1	48±1	53±2	50±0	50±0
10000	16	64	47±1	47±1	51±3	50±0	50±0
10000	16	128	42±1	43±1	50±2	50±0	50±0
10000	20	64	45±1	45±1	49±2	50±0	50±0
10000	20	128	40±1	41±1	48±2	50±0	50±0
10000	24	64	43±1	43±1	50±0	50±0	50±0
10000	24	128	38±2	39±1	50±0	50±0	50±0
10000	28	64	42±1	43±1	50±0	50±0	50±0
10000	28	128	36±1	37±1	50±0	50±0	50±0
10000	32	64	41±1	41±1	50±0	50±0	50±0
10000	32	128	35±1	36±1	51±0	50±0	50±0

Table B.3: Parameters used for training the federated model (number of participating clients n, batch size b, dataset used, and total data $|\mathcal{D}|$). We present the mean accuracy and the 95% confidence intervals for the mean (CI).

CIFAR-10 Results

$\|\mathcal{D}\|$	n	b	Training $\mu \pm$ CI	Testing $\mu \pm$ CI	Snapshot $\mu \pm$ CI	Distance $\mu \pm$ CI	Vanilla $\mu \pm$ CI
16000	2	64	62±1	60±1	64±1	49±1	51±0
16000	2	128	59±3	57±3	63±1	52±2	51±0
16000	8	64	57±1	56±1	60±1	50±0	50±0
16000	8	128	53±1	52±1	56±2	50±0	50±0
16000	16	64	51±1	50±1	54±2	50±0	50±0
16000	16	128	47±1	47±1	51±2	50±0	50±0
32000	2	64	66±2	63±1	66±0	48±3	52±0
32000	2	128	64±1	61±1	65±1	51±1	51±0
32000	8	64	60±2	58±2	63±1	50±1	51±0
32000	8	128	56±1	55±1	61±1	50±0	51±0
32000	16	64	56±1	56±1	61±2	50±0	50±0
32000	16	128	52±1	52±1	56±2	50±0	50±0

Table B.4: Parameters used for training the federated model (number of participating clients n, batch size b, dataset used, and total data $|\mathcal{D}|$). We present the mean accuracy and the 95% confidence intervals for the mean (CI).

	Snapshot		Distance		Vanilla	
n	Reg.	Self	Reg.	Self	Reg.	Self
2	66±0	68±0	48±2	55±0	51±0	52±0
4	65±1	63±1	50±1	51±1	51±0	51±0
8	62±1	63±1	50±0	50±1	51±0	51±0
16	59±1	61±1	50±0	50±0	50±0	51±1
20	55±2	59±2	50±0	50±0	50±0	50±1
24	52±3	59±2	50±0	50±0	50±0	51±1
28	54±3	60±1	50±0	50±0	50±0	51±1

Table B.5: Target models were trained on CIFAR-10 for varying participating clients (n) and total training data 24000. Comparison of actual attack accuracy to the estimated accuracy that adversaries can compute. Self-attack accuracy is computed by executing the attack on their own training data.

Appendix C

Interview Study Additional Materials

C.1 Additional Table

We include Table C.1 for reference. The table consists of examples from participants at the start of the study that they thought could be settings for private computation.

C.2 Interview Guide

Note that the order of the terms (a-h), the four scenarios (wage equity, census data, ad conversion, contact discovery), the four cases (one to four), and the examples within each case (a to do) were randomized.

C.2.1 Welcome

Welcome. Today we are going to be talking about a topic that may be new to you. We're currently studying public sentiments and understanding of novel data science techniques. We're interested in learning about what people expect and what questions they want addressed if their data is being used for data science by a company. The interview process helps us to understand these expectations and based on them, to make design recommendations for other researchers and policy makers. Please let us know at any point if you have questions. Before we start, I just want to make sure you have a something to write with/on, pen and paper. Throughout the interview, we're going to go through four types

Example data	**Private Data**	**Public Output**
(P1) Individual income, education completed	Individuals' incomes	Mean income by education
(P2) Voting	Individuals' votes	Result counts
(P3) Research study	Participants	Study data
(P5) Voting	Individuals' votes	Eligible voters
(P6) Income, location	Households' income	Mean income in a region
(P7) Salaries	Individuals' salary	Average salary
(P9) Financial organizations' data	Customer data	Financial trends
(P10) Telescope data	Raw data	Post-processed data
(P12) Personal data	i.e. age, demographics	Averages
(P13) Netflix views	Viewer distributions	Report on top service
(P17) Salaries	Individuals' salary	Average salary
(P18) Political surveys	Individual responses	Aggregated conclusions
(P19) Profits	Beneficiaries	Donations
(P21) Elections	Individuals' responses	Poll numbers
(P21) Infection disease studies	Collected data	Results

Table C.1: This table includes examples provided by participants in response to the prompt for "an example of a computation where the result can be made public, but the numbers used to determine the result are sensitive and need to stay private". Only responses that participants did not change their minds about are included.

of questions, some general, some about terminology, some about types of data sharing, and some about explaining how data is used.

On average I expect this interview to take 60 minutes. Do you have any questions or concerns before we start?

C.2.2 Warm-up/Baseline questions

To get us started, I'm going to ask you a general question on the topic. For the question, just state as many answers as come to mind and let me know when you're done.

1. Please list some of the ways that you expect companies use data about you and others.

C.2.3 Terms

For the next section of this interview, we are going to talk about approaches to data sharing that focus on 'how' the data is shared. We are going to go through a series of terms and I'll ask you if you are familiar with them, and some follow up questions.

1. Terms:
   (a) Private Computation
   (b) Encryption
   (c) Hashing
   (d) Multi-party Computation
   (e) Differential Privacy
   (f) Federated Learning
   (g) Private Machine Learning
   (h) Secure Computation
2. Have you come across the term [**(a) through (h)**] before?
   (a) (if yes) Where have you come across the term before?

(b) (if yes) What kind of guarantees do you think it provides to individuals? Some examples?

(c) (if yes) What do you think the purpose or goal is for a company using this?

(d) Please try to define the term in your own words

C.2.4 Describing Private Computation

We're now going to introduce the term private computation.

- A **computation** is just a calculation (generally in math). For instance, determining the largest number from a list, determining the average, determining a sum.
- A **private computation**, is a computation that tries to limit the information revealed by the result. It attempts to perform a computation (such as an average, sum, max), and share the result without anyone learning the values used to find the result.

1. What do you think is an example of a computation where the result can be made public, but the numbers used to determine the result are sensitive and need to stay private? Follow up: what is sensitive and what is not in the example.
2. How would you describe private computation in your own words?

C.2.5 Private Computation Scenarios

We are now going to talk about some different ways companies can work with client data.

I. Wage equity: An organization aims to identify salary inequities across demographics. They reach out to individuals and employment organizations about their salary data. The organization conducts an analysis over the salary data and produces a report on salary inequities. The organization acquires the data for the analysis such that... How acceptable is the organization's' goal? Scale: (completely unacc, unacceptable, neutral, acceptable, completely acceptable)

(a) ...salary data is shared directly. They receive the salary information of individuals from the individuals or employers via a web-based tool.

(b) ...salary data is submitted in a modified form privately (with technical and legal protections) via a web-based multi-party computation (MPC) tool. The technical protections prevent the identification of individuals' salary input from the final report. It also protects those who contributed their salary information from being connected to the salary information they provided (though does not prevent it from being known that they were a contributor). Using this technique can be more expensive for the analysis and they cannot use the data for any other purpose.

II. Census data is acquired from citizens of the country by the governing body. It includes information with respect to their age, gender, occupation, income, place of residence. The governing body analyses the data it acquires to inform policies and resource management. It can also make the results of the census available to researchers or the public by... How acceptable is the organization's' goal? Scale: (completely unacc, unacceptable, neutral, acceptable, completely acceptable)

(a) allowing aggregate/statistical queries (e.g. averages, sums, etc.) over the original data.

(b) allowing any query, but restricting individuals making queries from performing queries that allow them to make inferences/learn more information than is permitted. This means that some questions cannot be answered by querying the data.

III. Ad conversion: An online ad company wants to determine whether ads shown to its users lead to sales in physical stores. They reach out to a credit card company, which has transaction data for physical stores to compute whether there are purchases connected to their ads. The two companies perform the computation such that... How acceptable is the organization's' goal? Scale: (completely unacc, unacceptable, neutral, acceptable, completely acceptable)

(a) ...they each share their datasets. The credit card company shares the purchase data in physical stores and the online company computes the correlation to online identities locations and online ad views.

(b) ...the credit card company shares a modified version of their records. The credit card company shares the modified data such that the online company can only identify the financial records that correspond to its users. That is, the information on the other credit card clients (that do not use the online service) is not

available to the online company. Using this technique can be more expensive for the company and they cannot use the data for any other purpose.

IV. Contact discovery: A social media app wants to connect users that are already contacts with one another. The social media app has a list of contact information (its users) and the new user has a list of contact information (their friends etc). The app wants to determine the common contacts between the new user and the existing app users (the intersection). Note that not all of the new users contacts may use the social media app and not all users of the app are contacts with the new user. The social media app can connect the new user to existing users by performing a computation such that... How acceptable is the organization's' goal? Scale: (completely unacc, unacceptable, neutral, acceptable, completely acceptable)

   (a) ...the new user shares all their personal contact information with the social media app.

   (b) ...the new user shares a modified version of their personal contact information. The new user shares the modified data such that the social media company can only identify the new users' contacts that already use the social media app. That is, the other contacts (who do not use the social media app) are not available to the social media app. Using this technique can be more expensive for the company and they cannot use the data for any other purpose.

For each of [A], [B], [C], and [D], the following were asked:

1. How acceptable is it if the company uses (a)? Explain (completely unacc, unacceptable, neutral, acceptable, completely acceptable)

2. How acceptable is it if the company uses (b)? Explain (completely unacc, unacceptable, neutral, acceptable, completely acceptable)

3. What differences do you expect there should be (if any) if a company chooses to use **(b)** instead of **(a)**...

   (a) in general?

   (b) in terms of how companies inform their clients that their data is being used?

   (c) in terms of what companies inform their clients about when their data is being used?

4. How feasible/possible do you think it is for a company to use **(b)** instead of **(a)**

5. How should a company be explaining the technique **(b)** to their clients if they use it?

C.2.6 Potential Information Revealed

Case 1: One of the participating companies will additionally *be able to learn which specific records in the computed result correspond to you. How acceptable is it if the records that correspond to you are...*

a) ... your salary information? Explain.

b) ...your credit history (e.g., credit score, mortgage status)? Explain.

c) ...your location history (e.g., coordinates corresponding to your home, place of employment, etc.) Explain.

d) ...your genetic markers (e.g., for heart disease, cancer, etc.)? Explain.

Case 2: One of the participating companies will additionally *be able to learn if records of you were used to perform the computation. How acceptable is it if the records they learn correspond to you are in a dataset of...*

a) ...low-income households (and thus learn that you are in a low income household)? Explain.

b) ...dating app members (and thus learn that you use that dating app)? Explain.

c) ...people with a specific health condition e.g., diabetic, high-blood pressure, autoimmune diseases (and thus learn that you have that specific health condition)? Explain.

d) ...frequent drug users e.g., alcohol, marijuana, others (and thus learn that you are a frequent user of that drug)? Explain.

Case 3: One of the participating companies will learn *properties for groups*. A group could be people with glasses or any other attribute corresponding to a group of people such as demographics. How acceptable is it if a company can learn, for example...

a) ...glasses owners prefer shopping online? Explain.

b) ...women prefer shopping online? Explain.

c) ...glasses owners have poorer spending habits than non-glasses owners? Explain.

d) ...women have poorer spending habits than non-women? Explain.

Case 4: When two companies perform the private computation, if one of the participating companies possesses other additional information (e.g. statistics) *they can infer the exact value of a record used in the computation*. How acceptable is it if a company can always learn whether an exact record was contributed by the other organization? Explain.

a) How acceptable is it if a company can always learn whether an exact record was contributed by the other organization? Explain.

b) Is it more or less acceptable if a company can accurately learn the record contributed by a different company only 75% of the time? Explain.

c) ...50% of the time? Explain.

d) ...25% of the time? Explain.

e) To you, at what point (percentage) does this become unacceptable/acceptable? Explain.

Additional Information:

- How does it impact the acceptability if additional information has to be known to learn the values?
- How does the information that needs to be known influence the acceptability?
- How does the likelihood the additional information is known influence the acceptability?

C.2.7 General Responses

1. In general, how do you think companies should be communicating to their customers/clients about how they use customer/client data in general?
2. In general, how do you think companies should be communicating to their customers/clients about how they use customer/client data if they use private computation for the process?
3. In general, what do you think are companies responsibilities when using your data in these computations? Follow up depending on response: in terms of data protection responsibilities?

C.2.8 Participant Explanations

Prompt. The last thing we are going to do is an exercise called co-design. Even though you may have just learned about these techniques, we want you to think about how you would communicate these techniques to someone. There are no right or wrong answers. Imagine you work for a company that wants to use private computation. How would you communicate these practices to your clients? You can draw, write, verbally explain, etc.

Compare Show participant the previous suggestion.

- What would they add/remove to theirs based on it.
- What would they add/remove to the previous one.
- What is their final version they put forth after having considered the previous one.

C.2.9 Closing

Includes feedback and appreciation.