

# Enabling Retrospective Management of Data in the Cloud

BY

Mohammad Taha Khan  
B.S., Lahore University of Management Sciences, 2013

THESIS

Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy in Computer Science  
in the Graduate College of the  
University of Illinois at Chicago, 2020

Chicago, Illinois

Defense Committee:

Chris Kanich, Chair and Advisor  
Robert Sloan  
Ajay Kshemkalyani  
Blase Ur, University of Chicago  
Narseo Vallina-Rodriguez, IMDEA Networks Institute

Copyright by  
Mohammad Taha Khan  
2020

This thesis is dedicated to my grandparents *Anna* and *Nanu*, who passed away during my Ph.D.

## ACKNOWLEDGMENT

I am deeply grateful for my fantastic adviser Chris Kanich, who has continuously shown great support and encouragement throughout my Ph.D. Without his guidance, I would have never gone this far in this endeavor I took upon five years ago. Thanks, Chris, for being a great friend a making me feel welcome miles away from home. I would also like to thank Blase Ur and Elena Zheleva, for their guidance on this work. I am also grateful to the rest of my committee members Narseo Vallina Rodriguez, Robert Sloan, and Ajay Kshemkalyani for their generous time and help.

In addition, I am thankful to my undergraduate mentors Fareed Zaffar, Zartash Uzmi, Rashid Tahir, Matthew Caesar, and Momin Uppal. They provided me the opportunity to get involved in research and later on convinced me to attend graduate school.

I am also thankful to my graduate school collaborators Christopher Tran, Will Brackenbury, Shubham Singh, Noah Hirsch, Dimitri Vasilkov, Michael Tang, Maria Hyun, Miranda Wei, Mainack Mondal, Günce Su Yilmaz, Xuefeng Liu, William Wang, Cynthia Taylor, Joe DeBlasio, Geoffrey Voelker, Alex Snoeren, Sadia Afroz, Kyle Aguilar, Reem Hussein, Mohammad Abdul Salam, Phillipa Gill, William Tolley, Beau Kujath, Jedidiah Crandall, Zhou Li, Xiang Huo, Renata Alonso, and Joe Hummel. Even though that we did not do research together, I would like to acknowledge the friendship and support from the BITS Lab folks, Jason Polakis, Stephen Checkoway, Jakob Eriksson, Balajee Vamanan, Mohammad Ghasemisharif, Sara Amini, Peter Snyder, John Kristoff, Tomas Gerlich, Yanzi Jin, Timothy Merrifield, ABM Musa, and Sepideh

## ACKNOWLEDGMENT (Continued)

Roghanchi. Most of us were here for a good time, not a long time. I am also thankful to the wonderful administrative staff at UIC, especially Tricia and Sherice, for being so kind and understanding of any requests I made.

I am very grateful for my parents Seema and Iqbal, for their constant support and my brothers Talha and Talal, who have always looked up to me. I thank my friends Gulzar, Usman, Ghufan, Jerry, Rana, Myles, David, Huzaifa, Salman, and Haris, all who have been around with me to enjoy life during graduate school.

Finally, I am appreciative of the financial support provided by the National Science Foundation through grants CNS-1801644 and CNS-1351058, as well as the Open Technology Fund for their fellowship.

MTK

## CONTRIBUTIONS OF AUTHORS

**Chapter 1** provides an introduction to this work and **Chapter 2** is a summary of related works in the field. Content from these two chapters has been taken from certain pre-print versions of the following papers (1; 2). My adviser, Chris Kanich, Blase Ur, and Maria Hyun have contributed towards the writing of these chapters.

**Chapters 3** and **4** and from pre-print version of the paper Khan et al. “Forgotten But Not Gone: Identifying the Need for Longitudinal Data Management in Cloud Storage” which was accepted to the proceedings of *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2018. I was a joint co-author with Maria Hyun in this paper. My adviser, Chris Kanich, and Blase Ur were actively involved in the writing and research development process of this work. Maria Hyun, contributed to the writing of **Chapters 3** and the statistical analysis of the survey data in **Chapter 4**. Blase Ur and Maria Hyun also designed an initial draft of the survey in **Chapter 3** and produced the regression results in **Chapter 4**.

**Chapters 5, 6, 7, 8** and **Chapter 9** are from the pre-print of Khan et al. “Alethia: Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage.” for which I was the primary author. My adviser, Chris Kanich, Blase Ur, Christopher Tran, and Elena Zheleva have contributed towards the writing and research development of this work. In **Chapter 6**, I had help from Noah Hirsch, Dimitri Vasilkov and Michael Tang to transcribe the interviews. Will Brackenbury incremented our data collection framework with code that was used to extract text-based features discussed in **Chapter 7**. Shubham Singh assisted me with

## CONTRIBUTIONS OF AUTHORS (Continued)

the statistical analysis in **Chapter 8**. **Chapter 9** provides an overview of the design learning model using insights discussed in the subsequent chapters. I worked with Christopher Tran to set up a data-pipeline for training the model, as well as interpreting the results. Christopher Tran worked on the design of the classifier and produced the plots in this chapter. Both Christopher Tran and Elena Zheleva also contributed towards the writing of this chapter.

**Chapter 10** concludes the findings of this thesis and discusses potential future work. Content from this chapter has been taken from the pre-print versions of the papers (1; 2). My adviser, Chris Kanich, Blase Ur and Elena Zheleva have contributed towards the writing of this chapter.

## TABLE OF CONTENTS

<u>CHAPTER</u>		<u>PAGE</u>
<b>1</b>	<b>INTRODUCTION . . . . .</b>	<b>1</b>
<b>2</b>	<b>RELATED WORK . . . . .</b>	<b>6</b>
2.1	Cloud Storage Usage and Privacy . . . . .	6
2.2	Risks of Storing Information Online . . . . .	7
2.3	Retrospective Management of Online Data . . . . .	8
2.4	Understanding Data Sensitivity and Retention . . . . .	9
2.5	Interfaces and Automated Management for Online Data . . . . .	10
<b>3</b>	<b>EXPLORATORY STUDY METHODOLOGY . . . . .</b>	<b>12</b>
3.1	Overview . . . . .	12
3.2	Cloud Storage Services . . . . .	13
3.3	Data Collection . . . . .	14
3.4	Recruitment and Inclusion Criteria . . . . .	15
3.5	File Selection . . . . .	15
3.6	Survey Structure . . . . .	16
3.7	Data Analysis . . . . .	17
<b>4</b>	<b>EXPLORATORY STUDY EVALUATION . . . . .</b>	<b>20</b>
4.1	Overview . . . . .	20
4.2	Participants Demographics and Account Usage . . . . .	20
4.3	Account Archeology . . . . .	21
4.4	File Recognition . . . . .	22
4.5	File Management . . . . .	24
4.6	File Sharing . . . . .	29
4.7	Discussion . . . . .	31
4.8	Limitations . . . . .	33
<b>5</b>	<b>UNDERSTANDING SENSITIVITY AND USEFULNESS . . . . .</b>	<b>34</b>
5.1	Overview . . . . .	34
5.2	Approach . . . . .	35
<b>6</b>	<b>QUALITATIVE INTERVIEWS . . . . .</b>	<b>37</b>
6.1	Overview . . . . .	37
6.2	Methodology . . . . .	37
6.3	Perceptions Regarding Sensitivity . . . . .	39
6.4	Perceptions Regarding Usefulness . . . . .	40

## TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>		<u>PAGE</u>
<b>7</b>	<b>QUANTITATIVE STUDY METHODOLOGY</b> . . . . .	42
7.1	Overview . . . . .	42
7.2	Survey Flow . . . . .	42
7.3	File Feature Collection . . . . .	45
7.4	Ethics . . . . .	46
<b>8</b>	<b>QUANTITATIVE STUDY EVALUATION</b> . . . . .	47
8.1	Overview . . . . .	47
8.2	Demographics and Security Hygiene . . . . .	47
8.3	Categories of Sensitive and Useful Files . . . . .	48
8.3.1	Categories of Sensitive and Useful Files . . . . .	50
8.4	Management of Sensitive and Useful Files . . . . .	51
<b>9</b>	<b><i>ALETHEIA</i>: PREDICTING FILE MANAGEMENT DECISIONS</b> . . . . .	53
9.1	Overview . . . . .	53
9.2	Prediction Tasks and Baselines . . . . .	53
9.3	Dataset Description . . . . .	54
9.4	Experimental Setup . . . . .	55
9.5	Prediction Results . . . . .	56
9.6	Understanding Prediction Results . . . . .	59
<b>10</b>	<b>CONCLUSION</b> . . . . .	64
	<b>CITED LITERATURE</b> . . . . .	67
	<b>APPENDIX</b> . . . . .	73

## LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I	Stratified file selection categories for the exploratory study. . . . .	16
II	Participant demographics for the exploratory study. . . . .	21
III	Descriptive statistics of participants' Dropbox (DB) and Google Drive (GD) accounts. . . . .	22
IV	Broad scenarios used as prompts in our interviews. . . . .	38
V	File selection categories for the quantitative survey. . . . .	44
VI	A list of the features we automatically collected for each file using multiple APIs and custom code. . . . .	45
VII	Demographics for the 108 participants of the quantitative study (combined across rounds). . . . .	48
VIII	The percentage of participants who reported having files in categories implying they might be sensitive or useful. . . . .	49
IX	The percentage of files (N=3525) participants labeled as sensitive and useful, across different selection strategies. . . . .	50
X	Accuracy per file management decision and the incorrect (predictions). . . . .	59
XI	Top features for prediction tasks. Italicized <i>keywords</i> are top terms identified via the bag of words collections. . . . .	60
XII	Appendix - Factors correlated with file recognition. . . . .	85
XIII	Appendix - Factors correlated with file remembrance. . . . .	86
XIV	Appendix - Factors correlated with preferences for file deletion. . . . .	87
XV	Appendix - Factors correlated with preferences for file encryption. . . . .	88
XVI	Appendix - Factors correlated with wanting to stop sharing. . . . .	89

## LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	An design overview of the exploratory study survey. . . . .	12
2	File detail preview in file-specific section of our survey. . . . .	17
3	Comparison of file ownership and remembrance. . . . .	23
4	Participants' management decisions across the combinations of file recognition and remembrance. . . . .	25
5	Management decisions by participants for shown files. . . . .	25
6	Comparison of deletion and file ownership levels. . . . .	27
7	Comparison of of file encryption and the participants' technical back- ground. . . . .	28
8	Participants' preferences for file sharing. . . . .	30
9	Hypothesized file management decisions based on its usefulness and sensitivity. . . . .	34
10	Our approach for the follow-up quantitative study. . . . .	35
11	An design overview of the quantitative study survey. . . . .	43
12	The distribution of sensitivity and usefulness labels for files. . . . .	51
13	Desired file management by sensitivity/usefulness. . . . .	51
14	Precision vs. recall for predicting sensitivity and usefulness. . . . .	57
15	Comparison between directly predicting file management decision and first predicting sensitivity and usefulness for all files. . . . .	58
16	Predicted sensitivity probability for each document and image for every participant. . . . .	61
17	Preliminary classifier prediction precision as a function of predicted file sensitivity. . . . .	62
18	Appendix - The cumulative distribution of account age. . . . .	90
19	Appendix - The cumulative distribution of account size. . . . .	91
20	Appendix - The logarithmic distribution of account files. . . . .	92
21	Appendix - The cumulative distribution of shared files. . . . .	93
22	Appendix - Description of features collected from cloud files. . . . .	96

## LIST OF ABBREVIATIONS

2FA	2 Factor Authentication
API	Application Programming Interface
AUC	Area Under the Curve
FTC	Federal Trade Commission
GDLP	Google Data Loss Prevention
HITs	Human Intelligence Tasks
PII	Personally Identifiable Information
PIM	Personal Information Management
PRC	Precision-Recall Curve

## SUMMARY

Online cloud storage is a convenient and affordable way to store data over long periods. Today, millions of Internet users who have adopted cloud storage have accumulated years of information across thousands of files. The data stored is diverse, ranging from old photo-albums to old tax returns. As the social and personal contexts around this data continue to evolve, some of it loses its value and relevance over time. Even worse, some may contain sensitive information that puts users at risk. With the increasing prevalence of cyber crimes and data breaches, the decision to retain certain information online should be re-evaluated over time. Unfortunately, due to the scale of data, manual management of the cloud is infeasible, and there is a need for smart tools that allow users to achieve their desired management in an effective manner.

In this thesis, we present a comprehensive oversight into the problem. Through an exploratory study, we investigate the need for remediations for retrospective data management. Our results demonstrate a clear desire among users to manage files and the inability to do so due to the lack of practical tools. We then conduct a second study in which we carry out qualitative interviews to understand the kinds of management users intend. Finally, we incorporate the learned insights into the design of a learning-based tool (*Aletheia*). We predict users desired management decisions with an accuracy of 79%. *Aletheia's* performance validates a human-centric approach to developing management for files in the cloud. It also improves upon state of the art in minimizing the attack surface of cloud accounts.

# CHAPTER 1

## INTRODUCTION

**(Previously published as Khan, M. T., Hyun, M., Kanich, C., Ur, B. (2018). Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).)**

Cloud storage services are a convenient and affordable way to store data with full fidelity over long periods. Due to their increased popularity, major companies such as Google, Amazon, and Microsoft each have their version of a cloud storage solution (3). Even the free versions of these services provide gigabytes of storage, which is more than enough for thousands of documents and media files to pile up over the years. For instance, the default free version of Google storage is 15 GB for a user, and most users who do not have particularly large files, are likely to ever run out. As a result, many cloud based users across the globe have implicitly become long-term users of these platforms, with large amounts of their data accumulated in the cloud.

As the cloud becomes an integral part of the daily lives of Internet users, the data stored over time has become diverse, ranging from old tax returns and forgotten high school projects to pictures with exes and shared spring break photo albums. With the passage of time and evolving social contexts, some of the data loses its relevance. Crucially, some of it is no longer useful and may still be risky to continue to store.

This state of affairs has troubling consequences. While making indefinite retention of files the default option frees users from the risks of lost USB sticks or crashed hard drives; this policy also causes potentially sensitive information to accumulate in a single place. Such a form of data accumulation presents attackers with an attractive opportunity for account takeovers. If an attacker successfully impersonates the user (e.g., by guessing their password) or finds a flaw in the cloud implementation (4), they can access potentially all of the user's data. Moreover, the latent information stored within the content also puts individuals at higher potential risk in case of company-wide data breaches (5; 6), and lost devices (7). In addition to these potential privacy issues,

maintaining large amounts of data such that all of it is accessible on a moment's notice is a tremendous waste of resources. Just the sheer volume of data can lead to a highly unorganized state of the cloud account, which can cause the mixing up of valuable information with irrelevant data (8).

These concerns necessitate the need to reevaluate the data retention practices of users of cloud storage services. While prior research has characterized the need for retrospective data management in other domains (9; 10), effective management of cloud data is an open problem. Because revisiting thousands of files that have accumulated over the years is time-consuming, the foundation of any practical management approach must be some form of automated inference. Even so, the subjective and human-centered nature of file management requires an understanding of what makes a file in the cloud sensitive, as well as what makes it expendable. Overall, an ideal tool should incorporate learning mechanisms with user-centric management preferences to effectively increase the overall security and usability of consumer cloud services.

The projects discussed in this thesis focus on developing a comprehensive understanding of individuals' perceptions of the data stored in the cloud and creating learning-based tools, which allow users to manage this data efficiently. Our formal hypothesis is stated as follows:

*“We hypothesize that over the years, cloud storage services have evolved into sophisticated and versatile data-stores that contain information that is stale and even poses a privacy risk to users, this necessitates the development of methods that are specialized in accurately determining the extent of this risk and delivering precise retrospective remediation through automated management.”*

Through the process of empirical user studies, we first assessed the extent of sensitive and expendable data in the cloud, and after determining that the volume of data is infeasible for manual management, we took upon the task to explore users' interpretation on the kinds of management they intended, and integrated them into developing a learning based model for the protection and management of users' cloud accounts.

We first conducted an exploratory study to characterize the data stored in cloud accounts and investigated the need for remediations for retrospective data management. Through a survey of 100 Amazon Mechanical Turk participants, we asked questions about their perceptions of cloud storage and the files stored in them. A web application was developed, which integrated both the Google Drive and Dropbox Application Programming

Interface (API) to scan participants' accounts and collect the relevant file and user metadata. First, participants answered a set of generic questions. Next, they were shown a set of 10 stratified files from their cloud accounts. For each file, they specified if they would like to apply a specific management decision such as to delete, encrypt, and unshare the file (if it was shared), rather than keeping it in its current form. A sub-goal of the study was also to investigate how participants would feel about introducing semi-automated ways of retrospective management. To accomplish this, we asked participants questions regarding the ability to aggregate management decisions for files present in their cloud archives.

Results from the exploratory study demonstrated that participants had forgotten about the existence of a considerable amount of files in their accounts. A staggering 51% of the files shown were not remembered. In regards to file management, 83% of participants wanted to delete at least one file, despite reporting any storage shortage reasons. Overall, participants preferred to delete 34% of the files instead of keeping them as-is. Encryption was less popular, and participants only wanted to encrypt 7% of the files. Furthermore, they also preferred to unshare 11% of the files which they had previously shared with individuals. In addition, text-based responses about why participants wanted certain file management decisions were crucial in establishing the foundations of the follow-up study. Overall, our exploratory study provided evidence for the need for advanced file clustering techniques alongside understanding individualized user preferences to design learning-based solutions for management in cloud archives.

Qualitative insights from the first study highlighted users had strong preferences to manage and organize content in their cloud along the dimensions of sensitivity and usefulness. Due to the highly subjective nature of these concepts, as well as the incomplete understanding provided by prior work, we next explored users' mental models of these concepts qualitatively. To enumerate the many ways different people might think of a file as sensitive or useful, we conducted 17 qualitative interviews. We found that participants considered files sensitive for objective reasons like the presence of financial data or personally identifiable information, as well as subjective reasons like the presence of content deemed intimate to the participant's unique context. Participants considered files useful not only based on the recency of file access but also based on sentimentality and relationships.

Subsequently, we acted on this holistic understanding by constructing and evaluating classifiers through primarily a follow-up quantitative user study. A key challenge is that sensitive files, our primary target, are very much a minority class within cloud archives. As a result, for the quantitative study, we conducted two rounds of surveys. In each round, we showed participants dozens of files from their own Google Drive or Dropbox accounts, asking them to rate (and explain) the sensitivity and usefulness of each file. We collected numerous metadata and content features for each file, as well as for the cloud account overall.

In Round 1 we showed 75 participants files selected from their account using heuristics inspired by our qualitative interviews. We trained a preliminary classifier using the data collected. To further mitigate the class imbalance for file sensitivity, in Round 2 we showed 33 additional participants files selected based on our preliminary classifier. Using the combined data, we trained and evaluated a final classifier, which we dub *Aletheia*<sup>1</sup>.

The design of *Aletheia* was grounded on three prediction tasks. It predicted whether a user would perceive a given file as (i) sensitive and (ii) no longer useful. Finally, it predicted (iii) a file-management decision specifying whether the file should be kept, deleted, or protected (e.g., requiring 2 Factor Authentication (2FA)). To evaluate *Aletheia*, we used comparisons with the random and majority baselines. For the sensitivity task, we also used data features from the Google Data Loss Prevention (GDLP) API (11) as an additional baseline. *Aletheia* substantially improves state of the art for identifying files in the cloud that users are likely to perceive as sensitive and no longer useful. Predicting sensitivity, *Aletheia* showed an improvement of 52% for documents and 109% for images over the GDLP baseline. For predicting files as no longer useful, *Aletheia*'s Area Under the Curve (AUC) improvement was 51% for documents and 97% for images when compared to a random classifier. In regards to predicting participants' desired file management, the accuracy improvement was 49% over the most sensible baseline, a majority label classifier. We also noted that using scores of perceived sensitivity and usefulness increased the accuracy of management by 11%. This finding supports our approach to understand

---

<sup>1</sup>Aletheia is the Greek word for truth, which through the privative alpha literally means “un-forgetfulness” or “un-concealment.”

how users would interpret sensitivity and usefulness and incorporate it into the design process of automated management. To our knowledge, *Aletheia* is the first classifier that aims to identify files users for management decisions that combine both standard learning techniques and user-centered insights.

The rest of the thesis is structured as follows: In Chapter 2, we provide an overview of the prior work. Chapter 3 and 4 discuss the methodology and evaluation of the exploratory study. In Chapter 5, we detail our approach for the follow-up quantitative study. Chapter 6 elaborates our qualitative interviews, while Chapter 7 and 8 explain the methodology and results of our quantitative study. In Chapter 9 we discuss the design and performance of *Aletheia*, and conclude the thesis in Chapter 10.

## CHAPTER 2

### RELATED WORK

(Previously published as Khan, M. T., Hyun, M., Kanich, C., Ur, B. (2018). **Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage.** In **Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems** (pp. 1-12).)

Previous scholarship related to this project spans multiple sub-areas, including literature on the use of cloud storage, the risks and harms of online data, retrospective information management, user conceptualizations of online data attributes, and the development of tools and interfaces to manage online data.

#### 2.1 Cloud Storage Usage and Privacy

The advent of cloud storage was based on the reality of increasing amounts of data and decreasing costs for storage. The cloud provides more storage at a lower cost per customer, thanks to the efficiency of data centers. Cloud storage providers support both thick and thin client platforms (12) and ensure data availability, protected from failures (13; 14). As a result, cloud storage has gained significant popularity. Consumer cloud storage has developed primarily over the last decade. Box announced online file sharing for personal use in 2005, and Dropbox followed soon after. In 2019, the cloud storage market was valued at \$46.12 billion and is projected to reach \$222.25 billion by 2027 (15).

Despite its benefits, cloud storage has many implications for privacy and security. A careful analysis of the architecture and workloads of such systems has highlighted vulnerabilities in their usage, as well as how these issues impact users (13; 16). Computer experts have found security issues in the implementation of cloud storage. For example, Hu et al. (17) evaluated Mozy, Carbonite, Dropbox, and CrashPlan, finding that none offered any guarantees for data integrity and availability, nor assumed any liability for security breaches or data loss. Moreover, most free services did not offer data encryption, forcing data safety to become the user's

responsibility. Thus, when personal information is at risk, as in the 2014 case of Dropbox’s link disclosure vulnerability (18), users are left vulnerable. While legal protections on data stored in the cloud dictate that users do have a reasonable expectation of security and privacy in the cloud (19), the question remains: how do providers implement user-centered data management? To this end, our work explores this question in great detail, and we achieve it through a user-centric approach.

Issues around cloud privacy are worsened because users do not fully understand how their data is managed. It is not uncommon for private information to be uploaded to the cloud unintentionally; the majority of users in a study by Clark et al. discovered private photos in the cloud they did not realize were there (20). Although some solutions have been proposed to allow users to take advantage of the cloud without compromising privacy and autonomy (21), users still express distrust of the cloud. In Ion et al.’s cross-cultural study of cloud usage, most participants perceived cloud storage to be less secure than local storage (22). This would explain why users are reluctant to store sensitive data in the cloud (23; 24; 25; 26; 27). Many of these concerns could be mitigated if users had a better understanding of which files were stored in their cloud, as well as an active role in managing their data. Although researchers have analyzed user perceptions and system limitations, there has been little research from a user-centered perspective about what data users have stored in the cloud and forgotten about, as well as what they would like to do with that data. We take the first steps towards filling that gap. We investigate cloud storage usage, develop an understanding of why users initially store in the cloud and their desired file-management decisions.

## **2.2 Risks of Storing Information Online**

With the increasing use of cloud based platforms being used to retain longitudinal information, the possibility of being harmed by data leaks has become very real. The breaches can either happen at an individual level as a result of targeted attacks or on a larger scale, though orchestrated company-wide data breaches.

The Federal Trade Commission (FTC) (28) has raised the issue of how online data of individuals can be misused against them and can lead to financial, as well as health and safety risks. Similarly, academics have also studied (29; 30; 31) how the stealing and leakage of personal data by malicious actors can cause damage to

users of all age groups. Due to these inherent dangers, it is more than ever necessary to have better protection mechanisms for securing online data.

To better understand this risk, researchers have also taken an alternative angle by studying the privacy paradox. Kokolakis et al. (32) show how over the years, awareness around online privacy among users has increased. The variation of privacy perceptions has also been studied from a generational standpoint (33), from the perspective of the associated risks (34) in various scenarios, and the magnitude of various types of risks which include doxing (35; 36), stalking (37) and cyberbullying (38). Specifically, in the context of cloud storage, while Ion et al. (22) have determined that individuals understand the risks of storing data in the cloud, many users seek better privacy and are willing to pay for it.

Although sometimes beyond users' control, the ability to have user-centered management of data has become a crucial part of online security. A prime example of the lack of such a resource has led to various data breaches in the recent past. These include the 2012 Dropbox password leaks (5), the Ashley Madison breach (6), the 2019 Capital One hack (39) and the infamous Equifax data breach, which exposed sensitive data of millions of users and led to an FTC investigation that ended with a large settlement (40). For most data breaches, organizations are not able to recover the data or even provide fair compensation for the damage. While organizations have a stronger grounding and can recover, end users are the ones who end up taking a majority burden of these breaches. This looming risk to end users necessitates the need for active management of their data at an individual level. However, the scale of data makes management infeasible, and the question remains on what steps users need to take in order to secure the online ecosystem. Our work explores this open problem in great detail.

### **2.3 Retrospective Management of Online Data**

While surprisingly little work has investigated retrospective data management for cloud storage, extensive literature has examined similar questions for data on other online platforms. One significant sub-area is the management of social media posts to safeguard privacy. This is extraordinarily complex because users make dynamic privacy decisions based on context (41). Mondal et al. (9) demonstrate through a Facebook user study, the mismatch of privacy posts, and the desire among social media users to change the audience of their posts.

Although cloud storage and social media serve different purposes, this still provides a useful point of comparison, as both support content that can be either shared publicly, kept private, or removed altogether. Researchers have also developed Cloudsweeper, a cloud-based email protection system, which lets users remove or “lock-up” sensitive, unexpected, and rarely used information. While it effectively protects some sensitive files (10), Cloudsweeper’s methods do not map directly to cloud storage. More recently, a study performed by Brackenbury et al. (42) explores retrospective management based on file similarity metrics.

From research on retrospective management of social media, it becomes clear that a primary contributing factor for data management is temporality. It mediates whether users perceive content worthy of being acted upon, or left as-is, depending on its relevance over time (43; 44). The passage of time plays an important role, and users themselves cannot always predict what their preferences will be in the future (45). In a study (46) on Twitter, results show that even if users withdraw tweets (e.g., by deleting them), retweets may provide residual evidence and may even highlight when deleted tweets are missing.

Learning from this, we would expect that long term data in cloud storage can create similar problems for users. Decisions about managing files would depend heavily on the passage of time. Especially when sharing documents for the purpose of collaboration, one might expect temporality to influence the relevance of a document, and thus the associated management decisions. Determining whether users would want files to be deleted, encrypted, or archived is a complex calculation. Components ranging from file size to contents and access patterns, all need to be considered to provide efficient and useful file management. Our research contributes to specifying these considerations in the design of our file management methodology<sup>4</sup>.

## **2.4 Understanding Data Sensitivity and Retention**

A part of our work focuses on developing an understanding of data attributes, which make it more likely to manage from the users’ perspective. Related work in both academia and industry have explored this domain to better understand this concept.

A relevant management attribute in this regard is data sensitivity, which is subjective, and there is no universal definition of the term. Different organizations have come up with various categories of sensitivity. For

example, the Google privacy policy (47), mentions sensitivity associated with terms relating to race, religion, and sex. Similarly, Facebook and Twitter each have community guidelines that mention similar categories (48; 49). With the advent of cloud storage involving sensitive online data, Google had introduced its GDLP API (11) to classify and redact sensitive information from document files. The API categorizes numerous personal and financial identifiers as sensitive data. GDLP is primarily marketed to organizations and the healthcare industry to allow them to redact such information before storing data in the cloud. We use this API as one of our comparison baselines, however, our approach takes a much broader view of sensitivity to better capture consumers' conceptions of sensitivity.

In the academic domain, Peddinti et al. (50) have used anonymous Quora posts to explore contextual sensitivity. They identify categories of questions for which users are more likely to exercise anonymity. Vitale et al. (51) interview participants to understand data preservation categories of hoarding and minimalism, and also suggest alternative design strategies for personalized data management (52). Others have explored users' data deletion, retention, and coping strategies across complex cloud mechanisms, specifically when individuals have an incomplete understanding of what deletion in the cloud actually means (53; 54). Axtell et al. (55) have studied mental models of users across different age groups on practices regarding cloud storage usage. Similar research has looked at mental models of cloud data privacy for users with different technical backgrounds (56). Broader complementary work has evaluated file management practices by using only filesystem access information (57) and explored retention policies from an organizational perspective (58).

Learning from prior user-centric research, we begin our project with a qualitative study to characterize perceptions of data sensitivity and usefulness and incorporate this understanding in subsequent parts of our project.

## **2.5 Interfaces and Automated Management for Online Data**

The general research area of Personal Information Management (PIM) began in the 1980s to help users better store, organize, and retrieve collections of data (59; 60; 61; 62; 63). Researchers have suggested PIM interfaces for various online components. Dumais et al. (64) have developed a system of information retrieval that facilitates information reuse, while others have incorporated activity theory into their management interfaces (65).

Similarly, a lot of work has also been accomplished for email management(60; 66; 67; 68; 10; 69), and management of local files (61; 70). With cloud storage being a relatively new concept, little work has focused on cloud based PIM due to dynamic nature of their design and the lack of consumer understanding of cloud storage operations.

In a similar spirit, researchers have also proposed interfaces that use learning-based mechanisms within PIM interfaces. For social networks, Fang and LeFevre propose a “privacy wizard” for automatic privacy setting inference (71), while Ghazinour et al. (72) use collaborative filtering to recommend privacy settings. There have also been efforts to build classifiers around user-level privacy scores (73) and privacy risk (74). Similar research focuses on inferring sensitive attributes and identity matching in online platforms (75; 76; 77; 78; 79). Some have also used classifiers to predict private vs. public content (80; 81), as well as permissions for mobile apps (82) and the management image files (83).

To the best of our knowledge, we are the first to develop a classifier for automated management of files on cloud storage, that is motivated by user studies and and takes user-inference of file sensitivity and usability ratings into the design of a PIM classifier.

## CHAPTER 3

### EXPLORATORY STUDY METHODOLOGY

(Previously published as Khan, M. T., Hyun, M., Kanich, C., Ur, B. (2018). *Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage*. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).)

#### 3.1 Overview

To first understand the needs and opportunities for helping users manage forgotten files in their cloud storage accounts, we performed an exploratory study. Our procedure combined a dynamic online survey taken by our participants, along with programmatic data collection of metadata for files their accounts. Due to their popularity and API availability, we chose to implement our survey instrument for both Dropbox and Google Drive. We detail more on each provider in Section 3.2 and continued to use the same cloud services for the followup study described in Chapter 7.

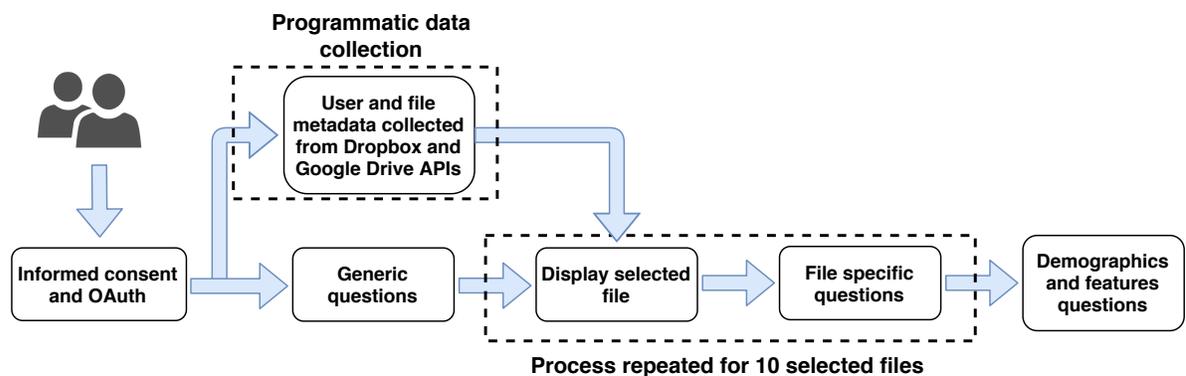


Figure 1: An design overview of the exploratory study survey.

Figure 1 summarizes our survey flow which has three main sections: one with a set of generic questions regarding the use of cloud storage, a second where we asked detailed questions about a stratified sample of ten files that each participant had in their actual Dropbox or Google Drive account, and a third in which we collected participant demographics and also asked about the potential for automating file management.

### **3.2 Cloud Storage Services**

Both Google Drive and Dropbox are similar at a high level; however some small differences impacted our study design<sup>1</sup>. Dropbox has existed since 2007, and Google Drive was introduced in 2012. Both offer free and paid tiers, Dropbox offers 2GB of free storage, while Google Drive provides 15GB, which is shared between all Google services, including Gmail and Google Photos. Dropbox and Google Drive provide sharing in two distinct ways. The first way of sharing files is to explicitly specify the recipient's account (or email address) in the cloud interface, done on an individual basis. The second method of sharing is to generate a link such that anyone with the link can access the file. Additionally, sharing can be transitive: a file shared from user A to user B can then be shared from user B to user C, depending upon the permissions given by user A.

At the time of the study, particularly Dropbox users intending to share an individual file could only give others view access, as granting edit access required sharing the entire folder containing the file. On the other hand, Google Drive allowed its users to grant view and edit access for both files and folders. Furthermore, for link sharing, Dropbox users with free accounts were limited to share links with view access only, whereas Google Drive permitted the apportion of view or edit access.<sup>2</sup>

Keeping these nuances into account, for our study we did not consider Dropbox files shared via link because they do not enable collaboration. In addition, we limited the selection of shared files where a user was not the owner. This was to prevent the selection of files from email attachments as a result of storage space shared among

---

<sup>1</sup>These services are dynamic and continue to change their operational strategies. Specifics reported in this thesis are subject to change in the future.

<sup>2</sup>This has changed as of March 2020. With the introduction of Dropbox Paper (84), users can now provide edit access to individual for both email and link sharing.

Google Drive and Gmail. We elaborate this further in Section 3.5 when explaining our file selection criteria.

### **3.3 Data Collection**

An essential part of our study involved showing participants files in their cloud storage accounts to ask questions that gauge their receptiveness to different data-management options. To achieve this, we first presented users with a consent form explaining what API access we needed and what information we would retain on our servers. After participants consented to the study, we requested access authorization to the service using OAuth2 (85), which allowed our application to programmatically access the files stored within the account. This granted us temporary access to these accounts without having to ask users for their passwords. We also provided guidelines in our privacy policy on how to revoke access once participants had completed the survey.

After obtaining authorization, we used the official API's provided by Dropbox and Google Drive to collect the data. Specifically, we used the Dropbox API v2 and Google Drive API v3. As the number of files per account varied widely and we needed the full list of files in the account to perform a stratified sample, we optimized API calls to ensure that the collection process was robust and relatively quick. As shown in Figure 1, we programmatically collected this data while the participant completed the generic portion of our survey.

Throughout this process, our primary concern was to maintain participants' privacy by collecting data in an ethical manner. We submitted our research protocol to the UIC Institutional Review Board (IRB) protocol number 2017-0186, and received an exemption. Details regarding the IRB are also present included the Appendix. We also used multiple techniques to protect user safety. First, we hosted our survey on an HTTPS domain with a valid certificate. We provided a detailed privacy policy with our contact details. For both cloud services, we limited the OAuth2 permission scope and requested only basic account information along with the file/folder metadata needed for our survey. We retained only the information we needed and stored one-way hashes for any unique identifiers to prevent retaining Personally Identifiable Information (PII). Furthermore, information such as file names and the names of other users who shared files with the participants were displayed in-browser via direct API calls and not retained on our servers.

### 3.4 Recruitment and Inclusion Criteria

For the exploratory study, we recruited participants on Amazon’s Mechanical Turk. We limited participants to North America and also required them to be age 18+ and have a previous approval rating of 95%+. As our goal was to investigate temporal file management and sharing decisions for cloud storage, we performed a preliminary screening of the survey participants using metadata from their accounts and verified that they met our criteria for inclusion, which we also presented to prospective participants in our Mechanical Turk Human Intelligence Tasks (HITs) description. Our criteria were:

- More than 50 total files on the cloud storage account
- At least one file that is older than 30 days
- At least 1 shared folder for Dropbox, and at least 10 shared files for Google Drive

These filters ensured that participants’ accounts were sufficiently well used for us to ask about various use cases. We also had additional sanity checks in our survey instrument to ensure that participants could not attempt to trivially meet our requirements without using their own legitimate accounts.

We recruited participants through two classes of HITs. In the first class, we asked participants to select the service (Dropbox or Google Drive) that they used more often for cloud storage. This resulted in 67 Google Drive users, yet only 17 Dropbox users. To even out this distribution, enabling us to compare more evenly across services, we posted additional Dropbox-only HITs, which resulted in an additional 16 Dropbox users.

### 3.5 File Selection

We asked each participant about ten different files from their cloud storage account. While random sampling of files would allow us to make statistical inferences about the entire contents of the cloud storage account, our focus was instead on collecting perceptions about as broad a set of files and use cases as possible. Thus, we conducted a stratified sampling strategy as outlined in Table I.

Within each of these ten categories, we randomly selected one file from all files that met the specified criteria. If no files in the user’s account matched a category (or if we had already asked about the only such file), we selected a random file from the account in its place.

<b>Index</b>	<b>Selected File Description</b>
1	Largest shared file of any type
2	Largest unshared file of any type
3	Shared media file of size greater than 250KB
4	Unshared media file of size greater than 250KB
5	Recently modified shared document
6	Recently modified unshared document
7	Old modified shared document
8	Old modified unshared document
9	Any shared file where participant is an editor
10	Any file shared via link (Google Drive only)

Table II: Stratified file selection categories for the exploratory study.

The first two categories (#1 & #2) are used to gauge perceptions of file size and sharing; we selected each of the largest shared and unshared files present in their cloud storage. Categories #3 - #8 select files by varying file types, recency of edits, and sharing status. Finally, to investigate how sharing modality affects answers, we varied the sharing modality for categories #9 and #10. As previously mentioned, because Dropbox users did not have the capability to share a file for editing via a link, category #10 on Dropbox was replaced with a file that satisfies category #3 instead. This stratified file selection enabled us to study various metrics across individual file types. After performing this study with 100 participants, we collected information about 1,000 files total. Due to an error, our survey software did not record three of these 1,000 responses. We thus report results for 997 total files in the exploratory study.

### **3.6 Survey Structure**

Our survey consisted of three main sections. We first asked participants about their usage of cloud storage and the general characteristics of their accounts. These questions covered attributes like the account age, primary reasons for using cloud storage, usage patterns, and account management.

The next section consisted of file-specific questions. Figure 2 shows a screenshot of what a participant saw at the beginning of each set of file-specific questions. Clicking the view button opened a new tab in the browser with a file preview provided by the cloud storage service. Participants were required to preview the file before

File 1/25

File Name: Passport.png

Location within Dropbox: /Document Scans/

You **MUST** view the file in Dropbox before proceeding.



View file in Dropbox

Figure 2: File detail preview in file-specific section of our survey.

they could proceed. This was followed by a set of questions about file recognition and remembrance. We also presented participants with three hypothetical file-management decisions: keeping the file as-is, deleting the file, and encrypting the file. We asked them to choose their preferred management decision. For shared files, we asked participants about the people with whom the file was shared with, and whether they would want to continue sharing the file with each of them.

Finally, we asked participants about their demographics, as well as about potential features that could be added to cloud-storage services. We collected basic demographics about participants, including age, gender, and profession. Among potential features, we asked whether auto-deletion, auto-archiving, and auto-encryption would be useful for the participants and, if so, in what circumstances. We include our complete survey instrument in the Appendix.

### 3.7 Data Analysis

After the data collection, we performed both quantitative and qualitative analyses. We provide the details of our evaluation methodology next.

**Aggregation and Basic Statistics:** Beyond survey responses, we also collected non-sensitive, non-personally identifiable metadata from participants' cloud storage accounts. Specifically, we calculated basic descriptive

account statistics, such as the number of bytes stored in the account, the number of files in the account, the types of files, and the percentage of files shared with others. We then aggregated this file metadata with our survey analysis, enabling more detailed insights.

**Qualitative Coding:** To analyze free-text responses, we followed a standard coding process. First, a researcher created a codebook based on the text responses. This codebook included labels for each response with definitions. After the first researcher finished creating the codebook, that researcher and another researcher read through the same survey responses and assigned a code to each using the codebook. After calibration on a small number of responses, both researchers independently coded all remaining participant answers and calculated the Cohen's Kappa coefficient to determine agreement on the coding. With a codebook that contained between three and fifteen themes per question, Cohen's Kappa between the two coders was at least 0.61 for each question.

**Regression Model:** To understand what file-level metadata, information about a given cloud storage account, and participant demographics correlated with participants' ability to recognize or remember files, as well as the decisions they made concerning managing the file and its sharing settings, we ran a series of mixed-effects logistic regressions. We chose a mixed-effects model because ten different files belonged to each participant, and our mixed-effects logistic regression models therefore include a participant-specific random factor to account for this non-independence of data. In each of our regression models, we included the following account-specific independent variables:

- service (Dropbox or Google Drive)
- age of the account (years)
- whether or not the account was used for work purposes
- whether or not the account was used for personal purposes

We also included the following file-specific factors:

- file type (document, image, spreadsheet, video, or other)
- access permissions (owner, editor, or viewer)
- number of days ( $\log_{10}$ ) since the file was last modified
- size of the file ( $\log_{10}$ )

- whether the file was shared, either with specific users or using a shared link

Because we hypothesized that usage patterns and management decisions might differ between Dropbox and Google Drive, we included terms to capture the interaction between the service and each of the five file-specific factors. In addition, we also included the following participant-specific factors:

- participant's age
- participant's technical background (defined as holding a degree or job in computer science or related fields)

We also ran an analogous ordinal regression to identify correlations between these factors and participants' preference about whether or not to keep sharing that file with up to three different individuals with whom that file was shared (sharing recipients). The dependent variable was ordinal, capturing preferences to keep sharing (1), whether it did not matter whether or not the file was shared (2), or to stop sharing (3). As this regression only included shared files, we removed the independent variable indicating whether or not the file was shared. However, we added an independent variable for participants' response about how recently they had been in touch with the sharing recipient (within the past year, over a year ago, or that they did not know who that person was). We treated both the participant and the file as random factors in our mixed-effects model. Because shared files were only a fraction of our data set, we did not include interaction terms. In Chapter 4 we report the p-values for factors that were significant. We provide the full regression tables in the Appendix.

## CHAPTER 4

### EXPLORATORY STUDY EVALUATION

(Previously published as Khan, M. T., Hyun, M., Kanich, C., Ur, B. (2018). **Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage.** In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).)

#### 4.1 Overview

We present comprehensive results from our survey data analysis, as well as significant highlights from our regression models. The goal of this analysis was to identify users' perceptions of file management in the cloud and understand the file specific factors that would help towards the desired file-management decisions.

#### 4.2 Participants Demographics and Account Usage

We begin by providing summary statistics for the demographics of our participants. The results are presented in Table II. Our participants were well-established users of cloud storage. In addition to using either Dropbox or Google Drive, 33% of participants also used Microsoft OneDrive, while 24% also used Apple iCloud.

A summary of the contents of participants' Dropbox and Google Drive accounts is shown in Table III. We also provide distribution plots of these account-level properties in the Appendix. While both services have been attracting significant numbers of new users in recent years (86; 87), our participants had been using these services for quite some time; 85% of participants' Google Drive accounts and 94% of their Dropbox accounts were more than three years old.

Participants used their accounts in several ways. Over 80% of participants used their accounts for both work/school and personal reasons, which can lead to an intermingling of files stored for different purposes with different sensitivities. Participants used their accounts frequently; 29% of participants said they use their account for work, school, or personal purposes at least once a week, and another 32% of participants reported

		Dropbox	GDrive
<b>Total # Participants</b>		33	67
<b>Gender</b>	Male	21	37
	Female	11	30
	Not answered	1	0
<b>Age</b>	<20	1	0
	20-35	18	47
	35-50	8	18
	51+	5	2
	Not answered	1	0
<b>Technical Background</b>	Yes	11	19
	No	21	48
	Not answered	1	0

Table III: Participant demographics for the exploratory study.

using their account at least once a month. It was relatively rare for the cloud to completely supplant local file storage, as 88% of participants reported retaining at least a subset of their cloud files on a local storage medium.

### 4.3 Account Archeology

Beyond analyzing usage trends, we also explored what types of files were stored on the cloud. Media files, which we defined to include sound files, images, and videos, had the most significant share (42%). We defined documents to include files with `.txt`, `.docx`, `.pdf`, and similar extensions. In total, 22% of such files. Documents were far more frequent than spreadsheets and presentations, which accounted for only 3% of the files. File extensions that did not fall in any of these categories were clustered as “other”, and these made up 31% of files. This other category included compressed archives, CD/DVD images, installers, and config files.

While participants’ self-reported responses suggested that they accessed their storage accounts frequently, we used file metadata to further investigate how often users edited (changed, rather than only viewed) files. The median number of days on which users modified at least one file was only 30 over a span of two years (730 days). This suggests that content modifications are likely to be performed by users on a particular day in

Property	Service	Min	Median	Max
<b>Account Age (Years)</b>	DB	0.4	4.9	8.2
	GD	0.1	4.9	5.3
<b>Account Size (GB)</b>	DB	0.1	2.0	54.1
	GD	<0.1	1.2	63.3
<b># of Files</b>	DB	53	514	66,604
	GD	59	424	22,163
<b>Shared Files (%)</b>	DP	<0.1	21.5	100.0
	GD	0.3	44.0	99.7

Table IV: Descriptive statistics of participants’ Dropbox (DB) and Google Drive (GD) accounts.

bulk, rather than on a daily basis. As we extracted this insight from the last modified date of a user’s file, the reported statistic is a lower bound because multiple edits to a single file would appear as only the most recent modification.

#### 4.4 File Recognition

After showing participants a file, we first asked whether they recognized the file (i.e., whether they knew what the file was after looking at it). We found that the vast majority of the files we asked about were recognized; only 10% of Dropbox files and 16% of Google Drive files were not recognized.

As described in the methodology, we ran a mixed-effects logistic regression to investigate what factors specific to the file, account, or participant correlated with whether participants recognized the files they were shown. Compared to the “other” file type, participants were more likely to recognize documents ( $p < .001$ ) and images ( $p = .027$ ). Unsurprisingly, compared to files for which they were the owner, participants were less likely to recognize files owned by others and for which they only had editor ( $p = .001$ ) or viewer ( $p = .011$ ) permissions. We observed a significant interaction effect in which participants were more likely to recognize files for which they had editor permissions if they used Dropbox, rather than Google Drive ( $p = .018$ ), but the cloud storage service otherwise did not significantly impact file recognition. We did not observe any significant correlations between

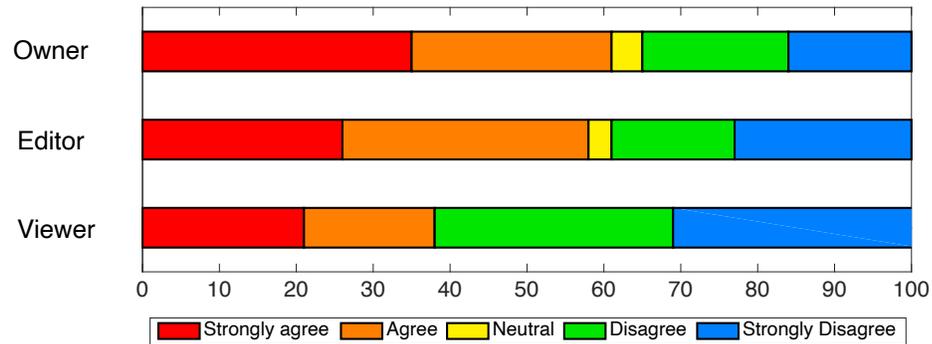


Figure 3: Comparison of file ownership and remembrance.

whether the participant recognized the file and any of the other file metadata factors or participant-specific factors we collected.

In addition to asking whether a participant recognized a file, we also asked whether they remembered that they still had that file in their cloud-storage account. Compared to simply recognizing the file, participants remembered retaining far fewer of those files: for 39% of Dropbox files and 34% of Google Drive files, they did not remember that the files were retained in cloud storage. While our non-random sampling approach is not representative of all files stored within these accounts, this result suggests that even though recalling the act of saving a file is not hard, with such large and long-lived accounts it is difficult to keep track of what has been retained.

Using a logistic regression, we found that compared to files in the “other” category, participants were more likely to remember video files ( $p = .025$ ), yet less likely to remember image files ( $p < .001$ ). Unsurprisingly, participants were less likely to remember files if they had only editor ( $p = .013$ ) or viewer ( $p < .001$ ) permissions, as opposed to being the owner of the file. Participants were also more likely to remember a file the more recently it had been modified ( $p < .001$ ) or the larger its file size ( $p < .001$ ). They were also more likely to remember shared files than unshared files ( $p < .001$ ). Participants were less likely to remember a file if their cloud storage account was older ( $p < .001$ ), although they were more likely to remember a file if they, the participant, were

older in age ( $p < .001$ ). Participants were less likely to remember files if they used their account for work purposes ( $p < .001$ ) and more likely to remember files if they used their account for personal purposes ( $p < .001$ ).

As shown in Figure 3, file ownership also had a positive correlation with remembrance. Through our regressions, we observed a number of significant interactions between file metadata and the cloud-storage service regarding file remembrance. Particularly, file ownership had a significant positive correlation with remembering the file was stored in the cloud ( $\chi^2(8, N = 862) = 32.244, p < .001$ ).

To investigate the utility of these stored files; we asked participants for a self-reported last accessed time for each file.<sup>1</sup> In the self-reported last accessed time, most files that we asked about had not been accessed recently. For Dropbox and Google Drive, respectively, 29% and 43% of files had last been accessed between one month and one year ago. An additional 41% of files had last been accessed between one year and five years ago. Regarding potential future utility, our participants answered that 30% of Google Drive files and 23% of Dropbox files would most likely never be accessed again. While copious cheap or free storage makes such “write-only” archives tenable, if a user were to store sensitive data there without expecting it to provide future benefit, the risks of such an archive clearly outweigh the rewards.

## 4.5 File Management

A key question we asked participants about each file was what file-management decision they would prefer for each file, chosen among keeping a file as-is, deleting it, or encrypting it in place. When asked about these capabilities in the abstract at the end of the study, participants had a more positive attitude about automatic encryption (72% agreed it would be helpful) than automatic deletion (32%). However, when asked about ten specific files, participants’ decisions were starkly different. Participants preferred that 58% of the files they saw be kept as-is, 35% files be deleted, and the remaining 7% of files be encrypted. These decisions are in line with participants’ self-reported priorities about file management overall; 40% of participants felt that never losing the ability to access files is important, while 26% of participants felt that protecting the file from unauthorized access

---

<sup>1</sup>Last access time, as opposed to the time of last modification, is not available via the API

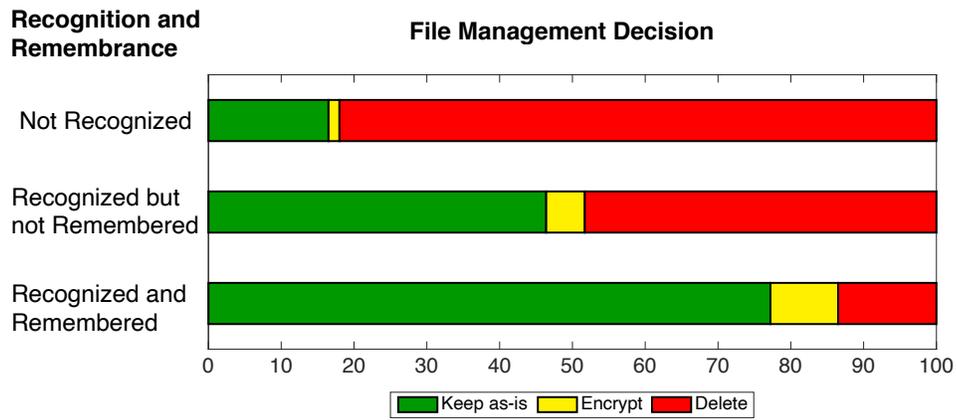


Figure 4: Participants' management decisions across the combinations of file recognition and remembrance.

is important. Figure 4 demonstrates how these management decisions were significantly correlated ( $\chi^2(4, N = 1000) = 260.26, p < .001$ ) with file recognition and remembrance.

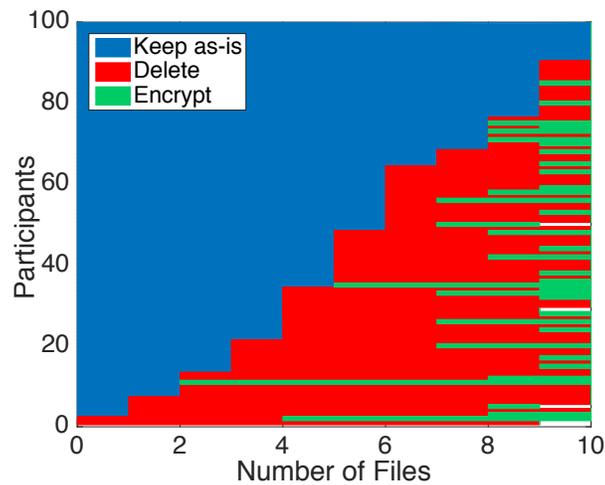


Figure 5: Management decisions by participants for shown files.

File-management decisions also varied across participants, as shown in Figure 5. While some participants preferred to keep everything as-is, 48% of participants wanted to delete or encrypt at least half of the files they

saw. Encryption decisions were motivated primarily by privacy. While some preferences for deleting files were also based on privacy-related concerns, decisions to delete a file were more commonly based on a file lacking future utility. Note that this tendency to delete files to clear useless clutter, rather than to maintain privacy, may be an artifact of the biases of participants who were readily willing to provide researchers access to their cloud storage accounts.

For each file shown, we asked participants to rate their agreement that it was important to prevent unauthorized access to the file. We used their responses to classify them into rough privacy personas. While four participants averaged at least “agree” to this statement across files (and could thus be considered privacy concerned), 25 participants averaged at least “disagree” (and could be considered marginally concerned), while the remaining 71 participants were in the middle (pragmatists). The file-management decision only varied for privacy concerned participants, who were more likely to encrypt than delete. Note that only four participants were in this category. Privacy concerned participants and pragmatists were more likely than marginally concerned participants to prefer unsharing currently shared files (unsharing 39%, 15%, and 3% of currently shared files, respectively).

We also asked participants whether their selected file management decision would apply to other files in their accounts, indicating that they could “describe those files using whatever language [they] use to think about them.” Responses were highly dependent on the specific files seen. Among decisions to keep a file as-is, 40% of participants indicated wanting to generalize such a decision to all other media files, while 30% wanted to generalize this decision to all files in their account. Among deletion decisions, 48% of participants wanted to apply such a decision to all other files they described as “not useful.” A common trend for generalizing the management decision was to apply it to similar file types, such as other ebooks or photos, as well as files contained in the same folder.

For 67 of the 100 participants, however, the file-management decision for at least one file would not necessarily generalize to other files. Among participants who would not want to generalize a deletion decision, 39% expressed a preference to examine deletion decisions individually. Other common reasons included not having other files of similar importance levels (36%) or not being aware of what the rest of their cloud storage contained (13%).

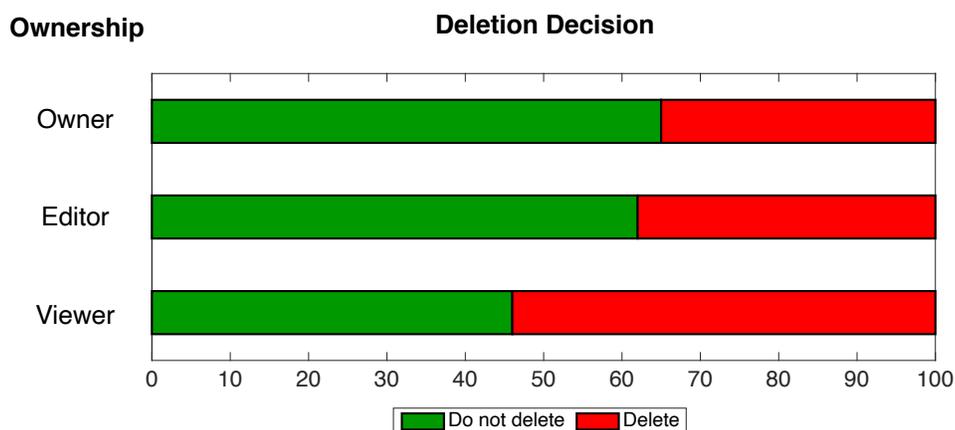


Figure 6: Comparison of deletion and file ownership levels.

Participants who chose not to generalize decisions to keep a file as-is stated similar reasons. In total, 30% of participants mentioned not having similar items or files of equal importance, while 24% said they preferred to examine files individually.

Participants were more likely to prefer deleting files if they had only editor ( $p = .008$ ) or viewer ( $p < .001$ ) permissions, as opposed to being the owner of the file. Figure 6 provides a more detailed comparison for this observation. These values were significantly correlated in our tests ( $\chi^2(2, N = 928) = 13.813, p < .005$ ). This effect, however, was far more muted for files on Dropbox than for those on Google Drive; there was a significant negative interaction between the service and the access permissions in predicting preferences for file deletion. We did not observe any other significant main effects, however, for predicting which files a participant would express a preference for deleting, nor which participants were more likely to delete files rather than keeping them as-is.

As with our regression model identifying which file-based, account-based, and participant-based features correlated with preferences for deleting a file, we observed few significant correlations between these factors and participants' preferences to encrypt a file. We observed that participants with a technical background, relative to participants without such a background, were more likely to choose to encrypt a file ( $p = .036$ ). Figure 7 depicts the correlation between participants' technical background and encryption decisions. Values in the figure were significantly correlated in our tests ( $\chi^2(1, N = 645) = 8.1447, p < .005$ ) Furthermore, participants who

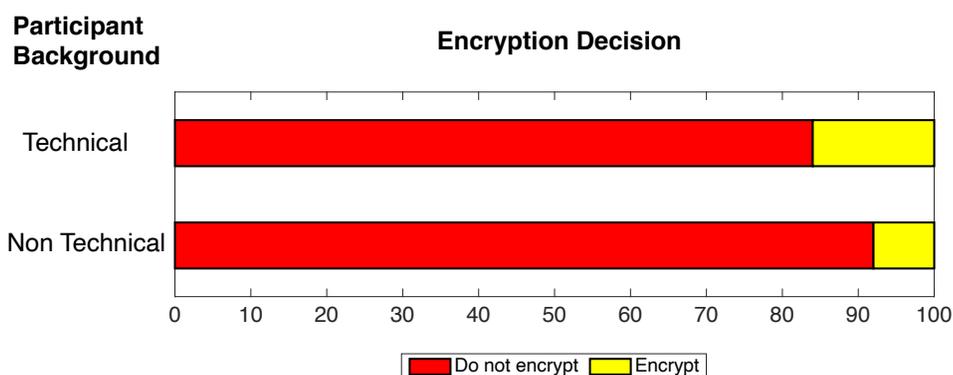


Figure 7: Comparison of of file encryption and the participants’ technical background.

used their cloud storage account for work purposes were less likely to choose to encrypt a file ( $p = .013$ ). We did not observe any other significant correlations.

Participants had multiple reasons for wanting to keep files as-is. When asked, 53% of participants said they might need the file in the future. One participant mentioned for a tax-related file, *“I might need it if I am ever audited, and I don’t know how long I need to keep tax-related [documents].”* In contrast, 38% of participants noted that files they would want to keep as-is did not contain private or sensitive information. For instance, one participant described, *“There is nothing about the file that I would be concerned about during a data breach.”* 28% of participants wanted to retain files for backup, while 26% mentioned they wanted easy access to the files remotely and across multiple devices.

For deletion, 91% of the participants who said they would like to delete at least one file mentioned the file was no longer useful or needed, or that it was causing clutter. For example, P27 explained, *“I don’t need [that photo] anymore and that folder is full of junk photos.”* 26% of participants said they chose to delete a file due to not being able to remember the file, and 10% of them mentioned deleting files to clear up space. Another popular reason for deletion was the file content being personal, with the goal of preventing unauthorized access. One participant mentioned, *“It’s a personal photo of my wife and I don’t want anyone else to see it.”*

While encryption was not as common as deletion, 65% of the 35 participants who encrypted at least one file stated securing against unauthorized access as their primary reason for choosing encryption. Commonly, participants’ responses suggested that these files contained sensitive information. For example, P44 mentioned

that one file *“is a financial document that I would not want to be public.”* We also observed instances where participants wanted to encrypt pictures and videos.

## 4.6 File Sharing

In addition to asking about preferred file-management decisions, we also asked whether participants wanted to keep sharing the files that were currently shared. We asked this question for the 212 shared files in our study. Since we asked about up to three other users with whom each file was shared, this resulted in 447 file-recipient pairs. We found that participants wanted to keep sharing with 41% of these file-recipient pairs, stop sharing with 11% of these file-recipient pairs, and did not have a preference for the remaining 48%.

In our regression of participants’ preferences about whether or not to continue sharing files that were shared with one or more other users by name, rather than through a shared link, we found that a handful of factors correlated with participants’ preferences. Unsurprisingly, participants were more likely to continue sharing a file when they had communicated in the past year with the recipient ( $p < .001$ ). In contrast, Dropbox participants were more likely to want to keep sharing files than Google Drive participants ( $p = .038$ ). Furthermore, participants were more likely to want to keep sharing when the file size was larger ( $p = .013$ ).

Whether participants were in touch (had communicated with the sharing recipient in the last year) was highly correlated with participants wanting to keep sharing files. Participants in touch with the recipient definitely wanted to keep sharing with the recipient for 59% of file-recipient pairs. In contrast, they definitely wanted to keep sharing for only 17% of file-recipient pairs when they were out of touch (had not communicated in the past year) and 12% of files where they did not know who the recipient was. Whereas participants definitely wanted to stop sharing for only 4% of pairs when they were in touch with the recipient, they definitely wanted to stop sharing for 23% of pairs where they were out of touch and 19% of pairs where they did not know who the recipient was. Figure 8 shows this distribution for our survey participants.

While the proportion of files participants definitely wanted to stop sharing with a particular person was similar for Dropbox (14%) and Google Drive (9%), the difference was in the strength of the preference to keep sharing. For particular file-recipient pairs, 57% of Dropbox participants definitely wanted to keep sharing the

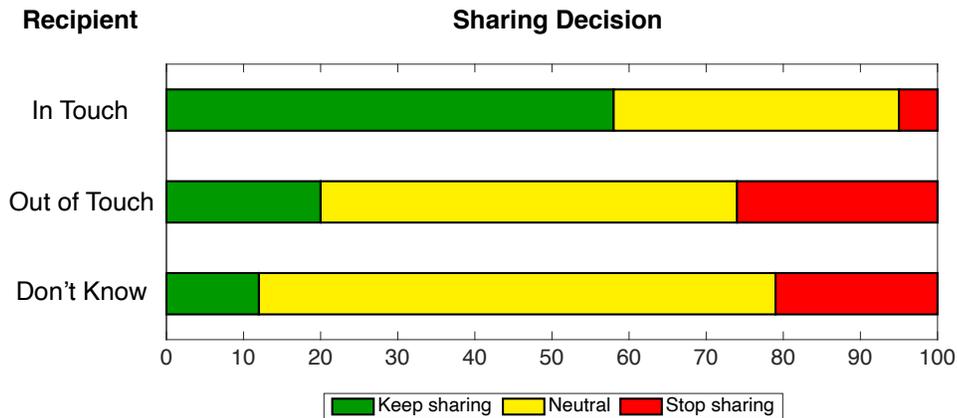


Figure 8: Participants' preferences for file sharing.

file. The same was true for only 22% of Google Drive file-recipient pairs. For the majority of Google Drive pairs (64%), participants did not care whether or not to keep sharing, whereas the same was true for only 34% of Dropbox pairs.

Whereas participants who used their account for work purposes did not care whether or not files continued to be shared for 53% of file-recipient pairs, the same was true for only 40% of pairs when participants did not use their accounts for work purposes. Participants who did not use their account for work purposes preferred at a higher rate either to definitely keep sharing or to definitely stop sharing files, compared to participants who did use their account for work purposes.

We also asked participants why they originally shared the file. The main reason was for work purposes (48% of responses). Other common reasons were to provide others access to a file (37% of responses), particularly media files, or to enable collaboration (11%). Participants who wanted to continue sharing the file gave similar reasons for originally sharing the file as those who did not. Furthermore, 17% of responses noted the file contained harmless information, while 3% noted that there was no reason to stop sharing the file. For example, one participant mentioned *“They don't need access to it for anything important, but it's not necessary to stop sharing.”*

On the other hand, participants also had a handful of reasons to stop sharing. 50% of responses mentioned that the task pertinent to the file had ended, while 39% of responses mentioned that the participant was no

longer in contact with the recipient of the sharing. Surprisingly, 11% of responses questioned why the file was being shared with that person. For example, one participant explained a decision to stop sharing by noting, *“I don’t remember sharing it with them in the first place.”*

Some files that remained shared with others after years of inactivity raises questions about whether users perceive these files as joint property, or whether they might prefer that long-shared files diverge into independent copies with time. This issue is exacerbated if the user has fallen out of touch with the other users with whom the file is shared.

We asked whether participants preferred that others’ edits be reflected in their copies of shared files, or whether they would prefer not to receive those edits for their own copy. For 61% of Dropbox files and 28% of Google Drive files, participants preferred to receive others’ edits. Conversely, for 51% of Dropbox files and 39% of Google Drive files, participants preferred that their own edits be reflected in others’ copies of the shared files. This decision was affected by whether the participant was an owner or editor of the file. For files owned by the participant, they preferred that their copies of the files reflect others’ changes 39% of the time, and that their changes be applied to others’ copies 52% of the time. For files with editing rather than ownership permissions, participants preferred that others’ changes be reflected in their copy 54% of the time, and that their changes be applied to others’ copies 44% of the time.

## 4.7 Discussion

Our participants had many files in the cloud that they had forgotten are there. When made aware of the existence of these files, the majority of participants wanted to delete, encrypt, or unshare at least one of the ten files they saw. Furthermore, participants did not even recognize 14% of the files they saw in the study, wanting to delete or encrypt 84% of these unrecognized files. These combined results highlight the need for retrospective file-management mechanisms in the cloud. Some retrospection tools already exist in other domains. For instance, Facebook has an “on this day” feature to highlight an old post, though this mechanism is focused on resharing. Whereas Facebook’s feature is meant to drive reminiscence and engagement, our results suggest that cloud users

also need such retrospective mechanisms to remind them of forgotten-about files, particularly those likely to arouse privacy concerns.

As a first step to understand automating the process of file management, we built regression models using basic file metadata and general information about the participant to try to predict file-management decisions. Unfortunately, these factors were not particularly strong predictors of users' file-management decisions. Because many of our participants had thousands of files stored in the cloud, simply encouraging users to manually revisit their files would present an undue burden. This exploratory study highlights why such an automated solution is challenging and requires a much more extensive understanding of the context of management as well as additional data collection.

In contrast, participants' free-text responses explaining how their decisions might generalize suggest that more advanced clustering of files, alongside identifying users' individualized preferences for managing files is needed. Participants wanted to manage files based on their underlying characteristics of sensitivity and usefulness. This suggests the need for machine learning approaches that use information extracted from the contents of files to perform an advanced clustering of useless and sensitive files. These specific insights motivate the approach of the next study discussed in Chapter 5 onwards.

Another prospective angle for future work is to combine techniques from machine learning with insights drawn from HCI work on users' security and privacy personas (88). Building on this stream of work, we imagine that users may naturally be categorized into different archetypes regarding their approaches to data management (e.g., those who favor deletion, those who prefer to keep sensitive files in disconnected storage, etc.). A predictive model could combine a deep understanding of the user's preferred mode of archive management with the specific management decisions already made for certain files. After the user makes a few representative file management decisions, these more advanced methods might be able to partially automate file management in order to ease the burden of retrospectively managing files in cloud storage.

To summarize, our findings highlight the disconnect between our participants' desired file-management decisions and the high overhead of retrospectively managing thousands of files in a cloud storage account. Overall results from this exploratory study provide evidence for the need for retrospective privacy mechanisms that em-

power users to manage the risks latent in their cloud archives without expending unreasonable effort.

## 4.8 Limitations

A core limitation of our study is that we report on a convenience sample. Our participants may not represent the typical user of cloud storage services, particularly since Mechanical Turk workers tend to be more technically oriented than the population at large. Furthermore, prospective participants with particularly sensitive files stored in the cloud might be reluctant to participate since they needed to give our software OAuth permissions to access their files. That said, even among individuals who were willing to participate, we observed many files participants would want to delete or encrypt.

Our study focused on Dropbox and Google Drive, which are only two of the many cloud storages services available, albeit the two most popular. We had an unequal distribution of Dropbox and Google Drive participants in our sample. A more comparably sized sample of the two services would provide a more accurate point of comparison.

While we included files generated by Google Docs, essentially Google’s online document-creation service, we could not include files from Dropbox Paper, a similar feature provided by Dropbox. An additional comparison of files generated by such web-based editing tools would have generated more comparable insights across the two cloud storage platforms.

By investigating our participants’ perspectives on a stratified sample of files stored in their own Google Drive or Dropbox account, we built a better understanding of the contents of cloud-storage accounts, identifying latent needs for retrospective file management tools. We used a stratified sample to measure a broad cross-section of files users retain in their cloud storage accounts, rather than focusing on the files most likely to arouse security and privacy concerns (e.g., files named “taxreturn2017.pdf” or that contain saved passwords). We take the limitations into consideration and ensure a more holistic approach to file selection and data collection from participants’ cloud accounts in our follow-up study.

## CHAPTER 5

### UNDERSTANDING SENSITIVITY AND USEFULNESS

#### 5.1 Overview

Insights from the exploratory study indicated that users had strong preferences to manage and organize content in their cloud archives along the dimensions of sensitivity and usefulness. We present our design approach for a study that aims to acquire a comprehensive understanding of these concepts from a user’s perspective and integrate these insights to develop a classifier for automating the file management process.

We envision specific types of file management based on the sensitivity and usefulness of files. Figure 9 shows our hypothesized management decisions for files in each of the four intersectional categories of sensitivity and usefulness. These decisions are based on prior qualitative insights. Specifically, we suggest that users of cloud storage are likely to delete a file if it is no longer useful, protect<sup>1</sup> useful and sensitive files, and keep regular files as-is. This management portfolio is essential to our study design and allows us to empirically evaluate our hypothesis.

---

<sup>1</sup>Protect was referred as encrypt in the exploratory study. We decided to update the terminology to enhance participant understanding.

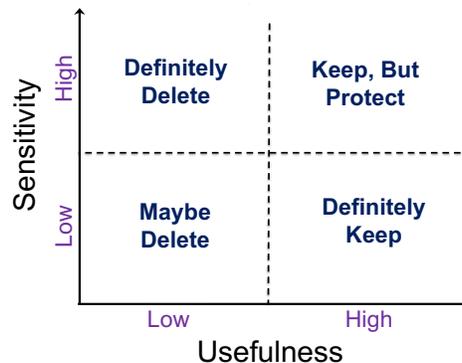


Figure 9: Hypothesized file management decisions based on its usefulness and sensitivity.

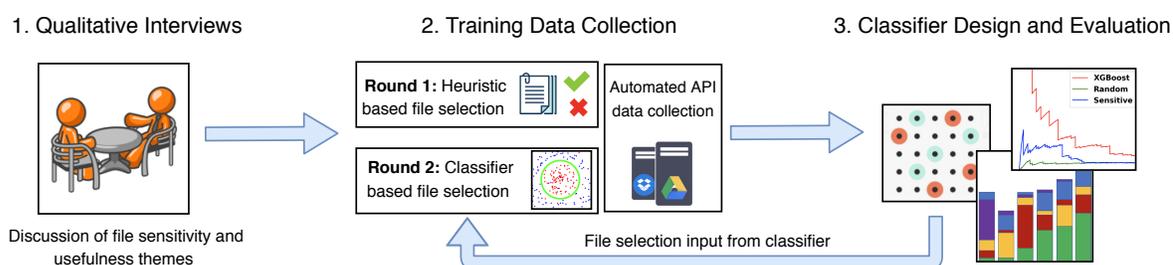


Figure 10: Our approach for the follow-up quantitative study.

Figure 10 shows a comprehensive overview of the design of our study, which we conduct in a multi-stage process. Due to the highly subjective nature of sensitivity and usefulness, as well as the incomplete understanding provided by the exploratory study, we first explore users’ mental models of these concepts qualitatively. We then transform the qualitative insights into realizable data attributes that can be collected from users’ cloud storage accounts in an automated fashion. Through a quantitative survey, we then collect labels and training data to develop and evaluate a classifier for automating the management decisions for files in the cloud. Similar to our exploratory study, we continue to use Dropbox and Google Drive as our cloud providers for this follow-up study. The next Section elaborates on our approach in further detail.

## 5.2 Approach

**Interviews Regarding File Sensitivity and Usefulness:** For files in the cloud, terms like sensitivity and usefulness can have subjective interpretations that vary across individuals. With the goal of enumerating the variety of these perceptions, we first conduct qualitative interviews. These interviews are in the form of open discussions to encourage individuals to highlight all possible file attributes associated with sensitivity and usefulness. Subsequently, we map these attributes to quantitative file features that can be collected programmatically. These interviews also influence the design of our primarily quantitative survey. Chapter 6 summarizes the findings for these interviews.

**Training Data Collection and Augmentation:** A prerequisite for developing an automated classifier is collecting training features and labels. The second phase of our study combines a user survey with the automated collection of various features about participants’ cloud accounts and files. These features include metadata provided by cloud storage providers, as well as deeper content analysis using third-party services like Google Cloud Vision (89). The survey is centered on showing participants files from their Google Drive or Dropbox accounts and asking them to label and explain their sensitivity and usefulness. We also collect file-management decision: whether they would want to keep, delete, or protect each file. As it is not feasible to show a participant all files on their account, selecting the right subset of files to yield a well-suited distribution of training data is a challenge. To solve this, we split our data collection into multiple rounds. In Round 1, we primarily use heuristic-based file selection using insights from our interviews. Because only a handful of files on a typical account are sensitive, heuristic-based file selection is likely to yield a small number of sensitive data points. Therefore, we train a preliminary classifier, using its predictions to select files in Round 2. Doing so lets us oversample the minority class (sensitive files). Chapters 7 and 8 further detail our methods and findings from both rounds of data collection.

**Developing *Aletheia*, An Automated Classifier:** Using the data collected from both rounds, we then build classifiers to predict file (i) sensitivity, (ii) usefulness, and (iii) desired management decisions. We formulate each prediction as a classification task. Note that file-management decisions are heavily influenced by file sensitivity and usefulness. As mentioned above, we use an initial version of the sensitivity classifier for Round 2 of data collection. Because decisions to delete data are highly subjective and consequential, we expect *Aletheia* to be used as part of a human-in-the-loop support system, rather than in a fully automated way. Therefore, we evaluate it with precision-recall analysis, which aligns with rankings of which files to present in a user interface or through recommendations. To quantify the accuracy of our models, we use the AUC metric for the precision-recall plots. Chapter 9 details *Aletheia*’s experimental setup and performance.

## CHAPTER 6

### QUALITATIVE INTERVIEWS

#### 6.1 Overview

To gain an initial understanding of how people conceive of the sensitivity and usefulness of files in the cloud, we conducted semi-structured interviews of cloud storage users. We aimed to build a formative understanding of factors that make someone perceive a file as sensitive or useful. This understanding underpins our quantitative study, eventually enabling us at scale to find files that may be sensitive, yet not useful.

We used this qualitative understanding both to develop closed-form survey questions and to identify meta-data and content features to collect about the files in participants' cloud accounts to train our classifiers. Chapter 7 lists these features and their relationship to these findings.

#### 6.2 Methodology

Using Craigslist, we recruited participants who had a Google Drive or a Dropbox account over three months old and were willing to attend an in-person interview. We interviewed 17 participants from January through June 2019. Our participants were local to Chicago and ranged in age from 20 to 45. To get a broader perspective, we prioritized participants without formal education or experience in an IT-related field. Six participants were full-time students, all from non-STEM majors. Compensation was a \$20 Amazon gift card.

Our protocol investigated participants' approaches to cloud storage, both abstractly and concretely (grounded in individual files in their accounts). The Appendix contains our complete interview script. This qualitative study protocol was submitted to the University of Chicago's Institutional Review Board (IRB) with protocol number IRB19-0098, and after review, received an exemption.

The first half of the interview focused on general reasons for using cloud storage, followed by open-ended discussion questions about the sensitivity and usefulness of broad classes of files stored in the cloud. We developed abstract conceptualizations of sensitive and useful files from these responses. To further spur participants'

---

<b>Scenarios for Sensitivity</b>
1. Files that would cause concern if they were hacked from the cloud
2. Cloud files that, if made public, would be embarrassing
3. Files that would cause worry if close family members viewed them

---

<b>Scenarios for Usefulness</b>
1. Files to be recovered if they were accidentally deleted from the cloud
2. Cloud files accessed and updated on a regular basis
3. Cloud files shared with friends and/or family

---

Table V: Broad scenarios used as prompts in our interviews.

thinking, we provided them the sensitivity and usefulness scenarios in Table IV. These specific scenarios were the research team’s initial hypotheses about how sensitivity and usefulness manifest.

The second half of the interview investigated the same phenomena more concretely. Participants logged into a web app we built that used the Google Drive and Dropbox APIs to show ten files randomly selected from their account. For each file, the participant explained its sensitivity and usefulness, giving us concrete examples of files that were sensitive or useful, in addition to specific attributes that made them so.

After the questions about particular files, participants were asked to provide overall feedback regarding draft questions from our quantitative survey discussed in 7. These specifically focused on various ways we might elicit perceptions of file sensitivity and usefulness. All interview responses were audio-recorded with consent and then transcribed using the Google Speech to Text API (90). As the goal of this study was to solely explore ideas and attributes of sensitivity and usefulness for cloud files, we did not collect any other information besides the interview recordings. A member of the research team open-coded these transcriptions to extract emergent themes. Then a second member of the team independently coded the transcripts using the resultant codebook. Finally, the two coders met and resolved conflicting codes.

### 6.3 Perceptions Regarding Sensitivity

In our general discussions of what makes a file sensitive, participants invoked the following seven classes of sensitivity:

**Personally Identifiable Information:** Files that contained names, contact details, dates of birth, passports, or driver’s license details were considered sensitive. Many participants cited their resume as an example. P01 explained, *“Anything that can easily identify you, like your name, your birthday, your phone number, your address. It’s all on my resume.”* P06 was concerned about identity theft, stating, *“I wouldn’t want to share publicly about my life. . . [as it makes people] targets of identity theft.”*

**Confidential Information:** Distinct from PII, participants mentioned that some data should never be released publicly because of its proprietary or confidential nature. Students mentioned original work that could be plagiarized. P05 said, *“If it’s like an essay or something that I’m turning in, I don’t think I necessarily want a bunch of people to read it.”* Three participants also mentioned files containing passwords.

**Financial Information:** Participants mentioned tax documents, pay stubs, and files with Social Security Numbers (SSN) as very sensitive. They also worried about statements for bank accounts and credit/debit cards, as well as other documents containing those numbers. Nine participants explicitly mentioned their SSN as particularly sensitive, yet also found on their cloud accounts due to backups of files like tax returns. P09 was especially concerned about *“God forbid, banking information.”*

**Intimate Content:** Participants described broad conceptions of content that could be considered intimate or personal, and thus sensitive. Photos, videos, and similar media files were most commonly mentioned, particularly individuals’ own photos (both in adult situations and in general), as well as adult content they had downloaded. P16 included among their embarrassing files *“Porn, anything that’s not for the public’s eyes. Pictures of myself or significant other.”*

**Personal Views:** Files that contained personal views or opinions were also identified as sensitive. P09 explained, *“I’m a religious person, and so there are times when I would make audio recordings or save videos that are of a religious nature. People may not particularly subscribe to it, or some people may deem offensive.”* Participants also mentioned files that contain political opinions and anti-government views.

**Self-Presentation:** Participants found files related to their self-presentation as sensitive. For example, P11 talked about *“unflattering photos and videos.”* Other participants said files that revealed activities they hoped to hide from specific people were sensitive. For example, P14 said, *“If there was a photo of me smoking weed, my parents would freak out.”*

**Content That May Be Misinterpreted:** Participants also said files that could be misconstrued by others were sensitive. A participant who was in the military discussed a specific picture they saw during the study by explaining, *“This is a picture of some of my soldiers at a cemetery. Even though it’s innocent, I don’t want people to associate this with, like, death.”* In contrast to data like financial documents, this type of sensitivity is particularly contextual and subjective.

#### 6.4 Perceptions Regarding Usefulness

Similar to sensitivity, participants also enumerated reasons why files were useful. Participants most commonly considered files in the cloud useful if they might need to access them in the future. The specific reasons for this future access spanned five categories:

**Reminiscence:** Participants frequently invoked photos’ sentimental nature and value for reminiscence as a key reason they are useful. P09 explained: *“Pictures are useful because they capture memories. You want to have some memory of good times, good events or different things.”* P16 explained why a specific picture of her kids was useful by saying, *“I would show my children what they looked like when they were younger.”* Similarly, P12 specifically invoked *“pictures of my children, husband, family experiences that we’ve had from travel, and pictures of lost ones.”* Expanding this definition, P09 explained, *“I share photos and videos of deceased family members that we like to reminisce about.”* Broadly, participants explained that files with sentimental value would likely remain useful forever.

**Active Projects:** Participants explained that files related to projects at work or school were useful, but many would not remain useful indefinitely. When asked to think about files they would prioritize recovering if accidentally deleted from the cloud, 13 participants mentioned work or school-related files. For example, P12

said, *“I would try to recover my resume and any school work that needed to be turned in.”* Similarly, P09 said, *“Documents are useful because... you always have to deal with documents online in school, at work.”*

**Recent Files For Reference:** Some documents remained useful for reasons other than their initial purpose. For example, P04 described a recent cover letter being useful for future job applications to additional employers by saying, *“This version is very current. I just recently updated it, so it will be very useful for me.”* In general, participants said files that had been recently accessed or modified were more likely to be useful, yet some older files also might be needed for reference.

**Files Frequently Updated Over Time:** Participants said cloud files that are frequently modified are useful. While some work- or school-related files fell into this category, journals and other evolving documents were key examples.

**Sharing:** Five participants mentioned that shared files were useful. For example, P03 (a student) explained: *“Midterm or final papers I usually store in the cloud if I need to share them with somebody else or have someone else look at it.”* This is an interesting insight as usefulness can not just be perceived individually.

## CHAPTER 7

### QUANTITATIVE STUDY METHODOLOGY

#### 7.1 Overview

Building on the robust insights from our qualitative interviews, we next conducted an online quantitative user study that combined an automated data collection process from participants’ cloud accounts. Our core goal was to collect rich data about participants’ perceptions alongside quantitative features of files in the cloud to train an automated tool for aiding cloud file management.

We first built a tool that allowed us to survey participants about specific files in their cloud storage account while simultaneously collecting metadata and content-based features about those files. As explained in Chapter 5, we collected data across two rounds. For each round, we recruited a separate set of participants to complete both the generic and file-specific surveys described below. In Round 1, we used a heuristic-based approach to select files. From the results of Round 1, we trained a preliminary classifier, which we used to select files in Round 2. We then used data from both rounds to build and evaluate *Aletheia*.

#### 7.2 Survey Flow

We recruited participants on Amazon’s Mechanical Turk and Prolific Academic. We recruited participants age 18+ from the North America region with a minimum platform approval rating of 95%+. Participants were also required to have a 3+ month old Google Drive or Dropbox account with at least a 100 files.

We first presented participants with a consent form about the study and a visualization of the data we would collect from their cloud account. Because of the sensitivity of this experiment, we took particular care in clearly explaining what data we were and were not collecting. We elaborate on the ethics of this further in Section 7.4. Afterward, we asked participants to authorize our tool to programmatically scan their account. Figure 11 summarizes the overall study flow and back-end data collection. The survey contained three sections: (i) generic questions about usage of cloud storage; (ii) file-specific questions about the sensitivity and usefulness of particular

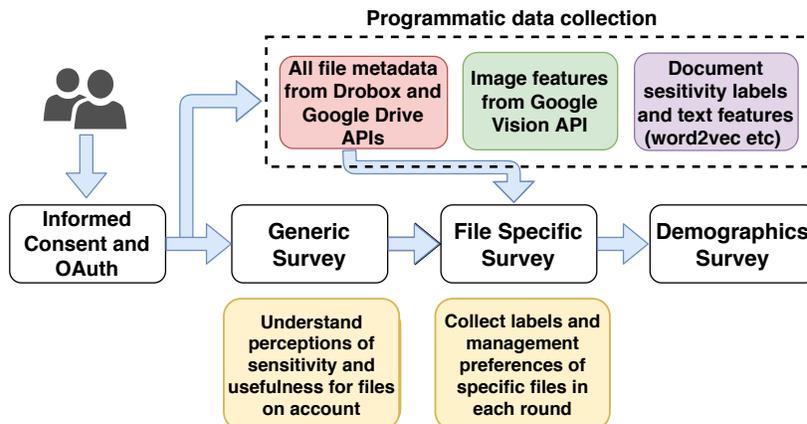


Figure 11: An design overview of the quantitative study survey.

files on their account; and (iii) questions about participant demographics and protection mechanisms used to secure their accounts.

The focus of our survey was its second part, in which we queried participants about particular files stored in their accounts. Participants’ responses, paired with the file features we collect, form the training data for *Aletheia*. As shown in Table V, our file-selection strategies differed across two rounds of data collection.

In Round 1, we selected files with heuristics defining different categories of files. For category #1, we looked for the presence of sensitive keywords in the filename. We chose keywords (e.g., “resume,” “passport,” “tax”) based on our interviews. The other three categories were documents (#2), media files (#3), and additional files (#4). We showed participants these files in randomized order. In comparison to a purely random selection across all the files in participants’ accounts, this approach provided a broader perspective, especially for accounts with a skewed file distribution (e.g., one with 10 documents and 500 images).

In Round 2, we used the data from Round 1 to train a preliminary classifier for identifying sensitive images and documents. Because sensitive files are a clear minority class (as most files are not sensitive), in Round 2, we used this classifier to select only potentially sensitive documents and images. In particular, we ranked all images (#5) and documents (#6) based on their predicted sensitivity score. We selected the top 25 images and documents, showing them in randomized order. While we had used just Amazon Mechanical Turk for data

Category	# of Files	File Description
<b>Round 1 (File selection based on heuristics)</b>		
1	5	Files containing a sensitive keyword in filename
2	8	Document files (txt, docx, pdf, xlsx, ppt etc.)
3	8	Media files (jpg, png, mp4, mpeg etc)
4	4	Files other than documents or media
<b>Round 2 (File selection based on preliminary classifier)</b>		
5	25	Top sensitive images
6	25	Top sensitive documents

Table VI: File selection categories for the quantitative survey.

collection in Round 1, we shifted over to Prolific in this round as it provided a more successful yield of participants due to its design catered for academic user studies.

For each file shown in either round, participants rated their agreement on a five-point Likert scale that “*I consider this file worth keeping,*”, which was the proxy we developed for usefulness based through our qualitative interviews. Similarly, an agreement that “*It would be risky, harmful, or otherwise dangerous if this file were accessed without my consent*” was our proxy for sensitivity. Our eventual goal was to train binary classifiers for finding files that are not useful, yet sensitive. Thus, we aggregated “strongly disagree” and “disagree” responses to the former statement as not useful, and “strongly agree” and “agree” responses to the latter as sensitive.

In addition to asking about the sensitivity and usefulness of the files, we also asked participants how they desired to manage the file. For management, they were provided three of the following options:

- **Keep as-is:** The file will remain in your cloud storage account in its current state.
- **Delete:** The file will be removed from your cloud storage account.
- **Protect:** The file will remain in your cloud storage account. However, you will need to take extra security steps to access the contents of the file.

Chapter 9 aims to predict the answers to the three management decisions. To better diagnose incorrect predictions, we also asked participants to justify each answer in free text. We provide our complete survey

Category	Collection Method	List of Features
Metadata	Google Drive/Dropbox API	account size, used space, file size, file type (img, doc, etc.), extension (jpg, txt, etc.), last modified date, last modifying user, access type (owner, editor, etc.), sharing status, sensitive filename
Images	Google Vision API	image object labels, adult, racy, medical, violent, logos, dominant RGB values, average RGB value
Documents	Local text processing	bag of words for top 100 content keywords, LDA topic models, TF-IDF vectors, word2vec representations, table schemas for spreadsheets
Sensitive Identifiers	GDLP API	<i>counts</i> of the following identifiers in a file: name, gender, ethnic group, address, email, date of birth, drivers license #, passport #, SSN, credit card, bank account #, VIN

Table VII: A list of the features we automatically collected for each file using multiple APIs and custom code.

instrument in the Appendix.

### 7.3 File Feature Collection

Table VI lists the features we collected. We chose these features primarily based on insights from the qualitative interviews. Because many interview participants mentioned personal and financial identifiers as sensitive, we used the GDLP API to find such identifiers in files. Likewise, because interview participants mentioned concerns about specific types of images, we used the Google Vision API to collect image object labels and binary labels corresponding to the presence or absence of adult, racy, medical and spoofed content within images. For docu-

ments, we performed local text processing to extract features, including TF-IDF vectors, topic models, word2vec vectors, and bags of words. Finally, we collected metadata about file activity and sharing.

## 7.4 Ethics

We obtained IRB approval prior to data collection. This was an extension approved under the protocol number 2017-0186 discussed in Section 3.3. Further details have also been included in the Appendix. Particularly, We took additional steps to protect participant privacy and ensure informed affirmative consent. Our consent page provided textual and visual examples of the type of data we collected about participants' files. Figure 22 in the Appendix provides a visualization of what participants during the consent process. We also provided a detailed privacy policy with our contact information. Our web apps were reviewed and verified by Google Drive and Dropbox, and our OAuth scopes were set precisely to those required for the survey. We did not retain any PII, and we only stored high-level labels, counts, features, and similarity-based hashes. We also guided participants on revoking access to our tool following completion of the study.

## CHAPTER 8

### QUANTITATIVE STUDY EVALUATION

#### 8.1 Overview

In our quantitative study, we had a total of 108 participants, 75 for Round 1 and 33 for Round 2. We collected free-text justifications alongside participants' Likert-scale perceptions of a file's sensitivity, its usefulness, and how the participant wished to manage the file. Thus, our dataset is rich with insights that we leverage in designing *Aletheia*. Except as noted, we aggregate results across both rounds of data collection because the distributions of responses were similar in most cases.

#### 8.2 Demographics and Security Hygiene

Table VII summarizes participant demographics. 78% of participants primarily used Google Drive, and 22% Dropbox. Participants were diverse in age and profession, which included engineers, technicians, freelancers, office assistants, salespeople, home-makers, and retailers. Participants were also well-established cloud storage users; 81% had used their account for 3 years or more. We observed both free and paid cloud accounts. Some participants used paid accounts provided by their work/school. All participants reported using their account for personal purposes, and 82% also used it for work/school. Participants were also reasonably frequent users of cloud storage; 22% of them used their account weekly, and 33% used it monthly.

Most participants were privacy-aware as over 50% of them reported that they would be moderately or extremely concerned if their cloud files were stolen in a data breach. While 43% had enabled 2FA, nearly one-fourth of participants reported taking additional steps to protect their accounts. These included using strong passwords, backing up information, and monitoring for malicious activity.

We asked our participants how they mainly access the files on their cloud accounts. 97% mentioned using a web browser, 90% used a smartphone app, and 61% used a PC-based application to synchronize files with the

Gender	Age		Technical Background		
Male	63	18–34	75	Yes	25
Female	44	35–50	29	No	82
Non-binary	1	51+	4	Not answered	1

Table VIII: Demographics for the 108 participants of the quantitative study (combined across rounds).

cloud. While these cloud applications initially grew out of PC-based synchronization tools, these participants are web and mobile-first users.

Characterizing account management was an important aspect of this survey. We asked participants to self-report how they felt about the organization status of their account. 47% of our them stated their account was not organized. We also asked how frequently they organized their account by deleting unnecessary files, moving files to different folders, or similar cleanup tasks. A majority of our participants did not perform any organization tasks on their accounts. Only 24% of our participants organized it once a month, or more frequently, 26% did it once a year, 24% performed it less than once a year, and 26% had never organized their cloud account.

### 8.3 Categories of Sensitive and Useful Files

In the general portion of the survey, we asked participants to provide specific examples in various categories of potentially sensitive or useful files. Table VIII summarizes these categories and the fraction of participants who reported that they had files belonging to that category in their account.

**Files Considered Sensitive:** More than half of participants stated that their account has files containing PII. Files in this category were related to bank accounts (20%), taxes (19%), their resume (11%), and IDs (11%). Discussing financial documents, one participant wrote, *“When I was buying a house, I might have uploaded some of the documents I needed for the mortgage onto the drive.”* While the presence of others’ PII was not very common (only 31%), such PII was typically that of school/work collaborators or family members. For example, P30 described *“tax returns that would have my family’s social security numbers and addresses.”*

Categories Implying Sensitivity	% of Participants
Files containing the participant’s PII	62%
Files containing PII of other than the participant	31%
Files with intimate or embarrassing content	30%
Files with original or creative content	84%
Files with proprietary information	23%
Categories Implying Usefulness	% of Participants
Files stored for future referencing	96%
Files with content of sentimental value	87%
Files which serve as backup	91%

Table IX: The percentage of participants who reported having files in categories implying they might be sensitive or useful.

For intimate and embarrassing content, all participants who had such files mentioned it being an image or video file. 76% of participants specifically referenced nudity or porn. In this regard, one participant explained, *“I have nude photos of my wife on there, and I might have some of myself.”*

Creative content was the most common category deemed sensitive. When asked about the specific type of creative work, participants mentioned school-related work (43%), art work (23%) and original writing (15%). Only 20% of participants expected that they had proprietary information in their account, of which 86% specifically identified it as work-related. For example, one participant wrote, *“There might be an NDA there but it is old and hopefully not any more of use.”* This sentiment and the mortgage document quote above exemplify the interplay between long-term archives and file sensitivity. While enabling long term storage is helpful, it can also accumulate sensitive files that are no longer even useful.

**Files Considered Useful:** Files that were in the cloud for future reference were the most common category of useful files, with 96% of participants mentioning such files. Common examples in this category were personal photos (21%), followed by documents for school (14%) and work (11%).

Among participants, 87% reported retaining files because of their sentimental value. For example, one participant wrote, *“I have a lot of my son’s first milestones, Christmas photos. I have photos of my wife and me*

Description (Selection Category #)	% Sensitive	% Useful
Sensitive keyword in filename (#1)	25%	65%
Media files (#2)	7%	67%
Document files (#3)	14%	61%
Other files (#4)	8%	51%
Top sensitive images from classifier (#5)	15%	66%
Top sensitive documents from classifier (#6)	15%	56%

Table X: The percentage of files (N=3525) participants labeled as sensitive and useful, across different selection strategies.

*before kids'. It helps me to remember how fast time flies.*". Another common category was videos and personal writings that belonged to the participant.

Files retained as backups were most likely to consist of many different file types. Common examples included images (21%), work (16%) and school documents (8%), and miscellaneous backup items (14%). Participants also mentioned files related to personal hobbies such as music, games, etc. For instance, one participant mentioned, *"I am a hobbyist musician, so I like to keep previous versions of songs I make on Drive. There have been occasions where I make a mistake later on and it's nice to have a previous version I can go back to."* Overall, 82% of files identified as sensitive or useful were images or documents. This fact, combined with the additional filetype-specific features of these files, led us to focus *Aletheia's* development specifically on these filetypes.

### 8.3.1 Categories of Sensitive and Useful Files

After we asked about useful and sensitive files in general, we showed each participant dozens of files from their accounts, asking them to label and explain their usefulness, sensitivity, and the desired management decision. This provided us with labels for a total of 3,525 files across both rounds. Among the files we selected (biased towards those that are sensitive), 62% were deemed useful, and 14% were deemed sensitive. Although the overall number of files perceived to be sensitive was low, 78% of our participants identified at least one file as sensitive. This observation aligns with previous studies showing the high overall likelihood of the existence of sensitive files stored in the cloud (20).

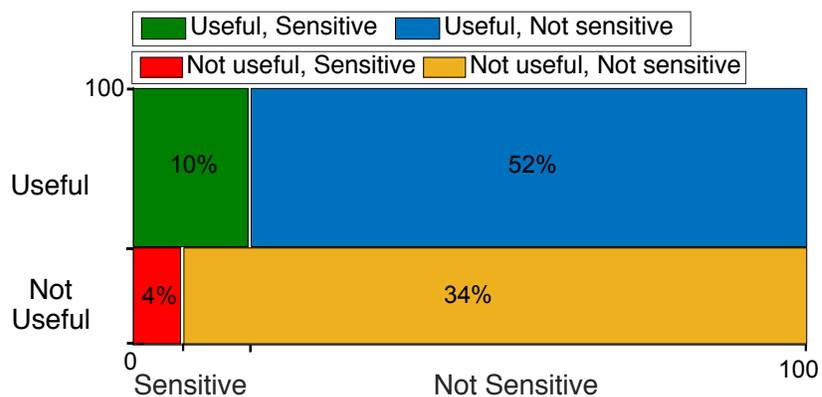


Figure 12: The distribution of sensitivity and usefulness labels for files.

Table IX summarizes perceived usefulness and sensitivity across the different file-selection categories. While documents and files with sensitive keywords in their filenames were more likely to be sensitive than media files and other file types, the distribution of file usefulness was fairly consistent across categories. Furthermore, Figure 12 is an area plot summarizing the distribution of files' usefulness and sensitivity. The percentages in each box represent the proportion of files belonging to each (sensitivity, usefulness) tuple.

#### 8.4 Management of Sensitive and Useful Files

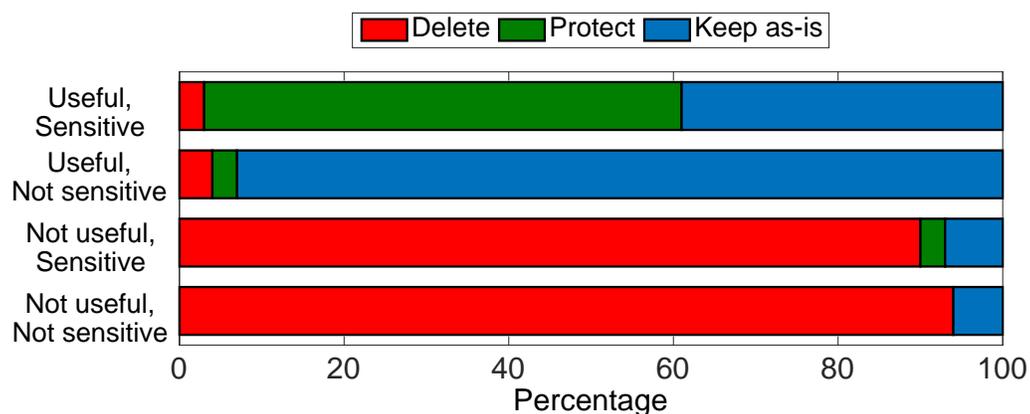


Figure 13: Desired file management by sensitivity/usefulness.

Revisiting our initial hypothesis regarding management, we see that Figure 13 directly aligns with the model we presented in Figure 9. For files deemed useful and not sensitive, participants wanted to keep 93% of such files as-is. For files that were not useful, the popular decision was to delete them, regardless of their sensitivity. This result is somewhat at odds with informal wisdom regarding digital packrats, but is in line with the proposed management decision. Namely, in 94% of cases, participants wanted to delete files that were not sensitive and not useful. When asked why they wanted to delete these files, the most common response was that they no longer needed them or that the files had served their purpose. The high likelihood of removing files shows both a willingness to reduce digital risk/clutter, as well as a lack of previous management that could have already deleted useless files from the account.

For files deemed sensitive and useful, participants wanted to protect 58% of them. Specifically, for Round 2, when files were selected using the classifier, we observed a 14% increase the protect percentage. In our model, we posited that users would be likely to protect all sensitive and useful files. Nonetheless, participants wanted to keep 39% of them as-is despite their sensitivity. This subjectivity in interpretations of sensitivity highlights the need for mechanisms that factor it into account through human input.

We also asked participants why they wanted to protect these files. Popular reasons included the file containing PII or financial information, files having sentimental value, and files containing intellectual property. Most of the reported reasons were consistent with the understanding of sensitivity we developed during the qualitative interviews. We observed a strong correlation between sentimental value and sensitivity. For instance, one participant wrote, *“This is a photo of a loved one I would like to keep private.”*

Overall, these management preferences and provided reasoning shed light on how participants conceptualize and operationalize file management in the cloud based on files’ perceived sensitivity and usefulness. Next, we leverage both our collected training data and these qualitative observations to build our *Aletheia*, which is an automated inference approach to predict file usefulness, sensitivity, and management decision. *Aletheia*’s ultimate goal is to assist users in protecting (or deleting) the files most likely to be in need of reconsideration.

## CHAPTER 9

### ***ALETHEIA*: PREDICTING FILE MANAGEMENT DECISIONS**

#### **9.1 Overview**

Users can have hundreds to thousands of files in their cloud storage, and our primary goal is to alleviate the burden of manual file management with automated tools. In this Chapter, we elaborate on the development of *Aletheia*. We explain the task of file management decision prediction based on automatically collected file and user account features. We also look into predicting user perceptions of usefulness and sensitivity to inform the file management decision classifier.

#### **9.2 Prediction Tasks and Baselines**

*Aletheia* has three prediction tasks: predicting whether a user will perceive a file as sensitive (task 1); predicting whether a user will perceive a file as no longer useful (task 2); and predicting what management decisions a user will take between keeping, deleting and protecting a file (task 3). To perform classification for each task, we compared several established supervised learning algorithms: Decision Trees (DT), Logistic Regression (LR), Random Forests (RF) using scikit-learn (91), XGBoost (XGB) (92), and Deep Neural Networks (DNN) with the Adam optimizer using scikit-learn. All model parameters were optimized using a grid search on the training set in each of the cross-validation folds and tested on the testing set. We used the best performing classifier, which turned out to be XGBoost for both the preliminary classifiers trained on Round 1 data and the final classifiers trained in Round 2. Here, we report results only on the final classifiers, which we refer to as *Aletheia w/ all features*, or *Aletheia* for short, together with three baseline models.

We defined the first baseline as a random classifier (*Random*), which randomly picked files for each management decision. Our second baseline was a majority classifier (*Majority*), where we always predicted the most frequent class. For the sensitive file prediction task, we employed a more meaningful third baseline, *GDLP feature count*, which leveraged the GDLP tool as described in Table VI.

This baseline used the output of the GDLP and ranked files based on the number of sensitive features in each. We also considered a variant of our approach, which instead of simple counts, took into account the output of the GDLP as features for predicting sensitivity, *Aletheia w/ only GDLP features*. We chose these baseline classifiers because, to the best of our knowledge, no prior work has attempted to predict perceptions and file management decisions for cloud storage files.

### 9.3 Dataset Description

For classification, we used the final dataset collected in Round 2. Our dataset consisted of tuples  $(X, Y)$ , where  $X_i$  is the feature vector, and  $Y_i$  is the target for prediction. The feature vector  $X_i$ , included metadata and information on files and user accounts. From user accounts, we had their usage statistics: how much storage they had and how much storage they used. For file information, we used the size of the file, whether the file was shared or not, the link access (view or edit), whether the file was last modified by the user or not, and the access type (owner, editor, viewer). For documents and images that contained text, we extracted counts of sensitive information discovered using GDLP API. In addition, we collected a bag of words on a heuristic set of keywords that may point out sensitivity. For documents, we collected an average word2vec embedding of each document using Google News Word2Vec embeddings (93). This approach enabled us to get textual context features without breaching the privacy of users and having the actual interpretable text from their files. For images, we used the Google Vision API to get multiple image features; this consisted of object labels in the image, as well as the presence of any adult, violent, medical, or racy content. We additionally converted the labels from the API output, including the “best guess label,” to one-hot encoding representation, as well a word2vec embedding representation, which were then added to the feature vector. A complete list of features is listed in Table VI. Compared to *Aletheia w/ only GDLP features*, which used the input from the GDLP API, we considered a more broad set of features that were both file-based and user-based.

For file management decisions (task 3), we used all files for which users answered questions. For task 1 and task 2, we broke down the evaluation by image files and document files since they have different features. The labels  $Y$  for each task were obtained from the user answers. Based on the answers to S-1, we considered

two labels for task 1, sensitive (“strongly agree”, “agree”) and not sensitive (“neutral”, “disagree”, “strongly disagree”). There were 15% of sensitive files. Based on the answers to U-1, we considered two labels for task 2, useful (“strongly agree”, “agree”, “neutral”) and not useful (“disagree”, “strongly disagree”). There were 38% of not useful files. Based on the answers to M-1, we had three labels for task 3, delete (40%), protect (8%), and keep as-is (52%).<sup>1</sup>

## 9.4 Experimental Setup

We describe our experimental setup for three different tasks. Tasks 1 and 2 for predicting sensitivity and usefulness had the same setup, while predicting file management decision in task 3 used a different setup.

**Task 1 and 2, Sensitivity and Usefulness:** We performed 5-fold cross-validation and report averaged results across 5 test folds. Since we focused on finding files that users wish to manage, we ordered examples in the test data by the probability of being  $Y_i = 1$  sensitive for task 1 or not useful for task 2, and assessed the precision and recall. This is a common setup for evaluating binary classification where one label (e.g., sensitive) is more important than the other (e.g., not sensitive).

We aimed for both high precision and high recall, but there is typically a tradeoff between them. Precision is computed as  $TP/(TP + FP)$ , where TP is the number of true positive examples (actual label positive, predicted label positive), and FP is the number of false-positive examples (actual label negative, predicted label positive). Recall is computed as  $TP/(TP + FN)$ , where FN is the number of false-negative examples (actual label, positive, predicted label negative). A Precision-Recall Curve (PRC) allowed us to see the tradeoff between predicting as positive a large number of files vs. a small number of files. For example, if we predict the top 20% of files, 20% corresponds to the recall and has a specific precision associated with it. We also use AUC, which specifies the area under the PRC and summarizes PRC as one number. A high PRC is generally desirable.

**Task 3, Management Decision:** In this task, we had three classes for classification: delete ( $Y_i = 1$ ), protect ( $Y_i = 2$ ), and keep as-is ( $Y_i = 3$ ). Since file perceptions of sensitivity and usefulness correlate highly with file

---

<sup>1</sup>S-1, U-1 and M-1 are question IDs of our survey in the Appendix

management decisions, we wanted to leverage them in our classification. However, we typically didn't have these labels for all files in a person's file storage. Therefore, we proposed to predict these labels using the classifiers for tasks 1 and 2. Then we could add the predicted labels as two additional features in the management decision prediction task. We compared the performance of adding these two features with a classifier that does not use them, a majority classifier, and an Oracle that has the actual perceived sensitivity and usefulness of a file. We performed an 80-20 training and testing split where all three classifiers had access only to the training data, and report accuracy on predicting file management decision on the testing set.

## 9.5 Prediction Results

We present the PRC averaged over five folds for sensitivity and usefulness, separated into image and document files. We also analyze the top features for predicting sensitivity and usefulness. For task 1 and task 2, a majority classifier performed the same as a random classifier for precision and recall, so we do not report results for random classifiers.

**Task 1, Sensitivity:** We studied whether it is possible to predict if a user will perceive an image or document as sensitive. Figures 14a–14b show the PRC for the sensitivity dataset. For documents *Aletheia* performs the best overall, while *Aletheia w/ only GDLP features* performs worse. The *GDLP feature count* classifier does not perform better than the *Random* baseline in this setting. *Aletheia* has an AUC 0.61, while *Aletheia w/ only GDLP features* has an AUC of 0.40, an improvement of 52%.

For images, *Aletheia* has an AUC of 0.71 compared to *Aletheia w/ only GDLP features* with an AUC of 0.34, an improvement of 109%. The *GDLP feature count* classifier has an AUC of 0.20. Compared to prediction for documents, we observe a much better performance in terms of the PRC for *Aletheia*, but not for *Aletheia w/ only GDLP features*.

From the sensitivity results, we show that a broader set of features besides sensitivity counts provides more accurate results. We also see *Aletheia* performs better at predicting sensitivity of images than documents. This makes sense, since we had additional image features on adult, racy, spoof, medical, and violent content, which

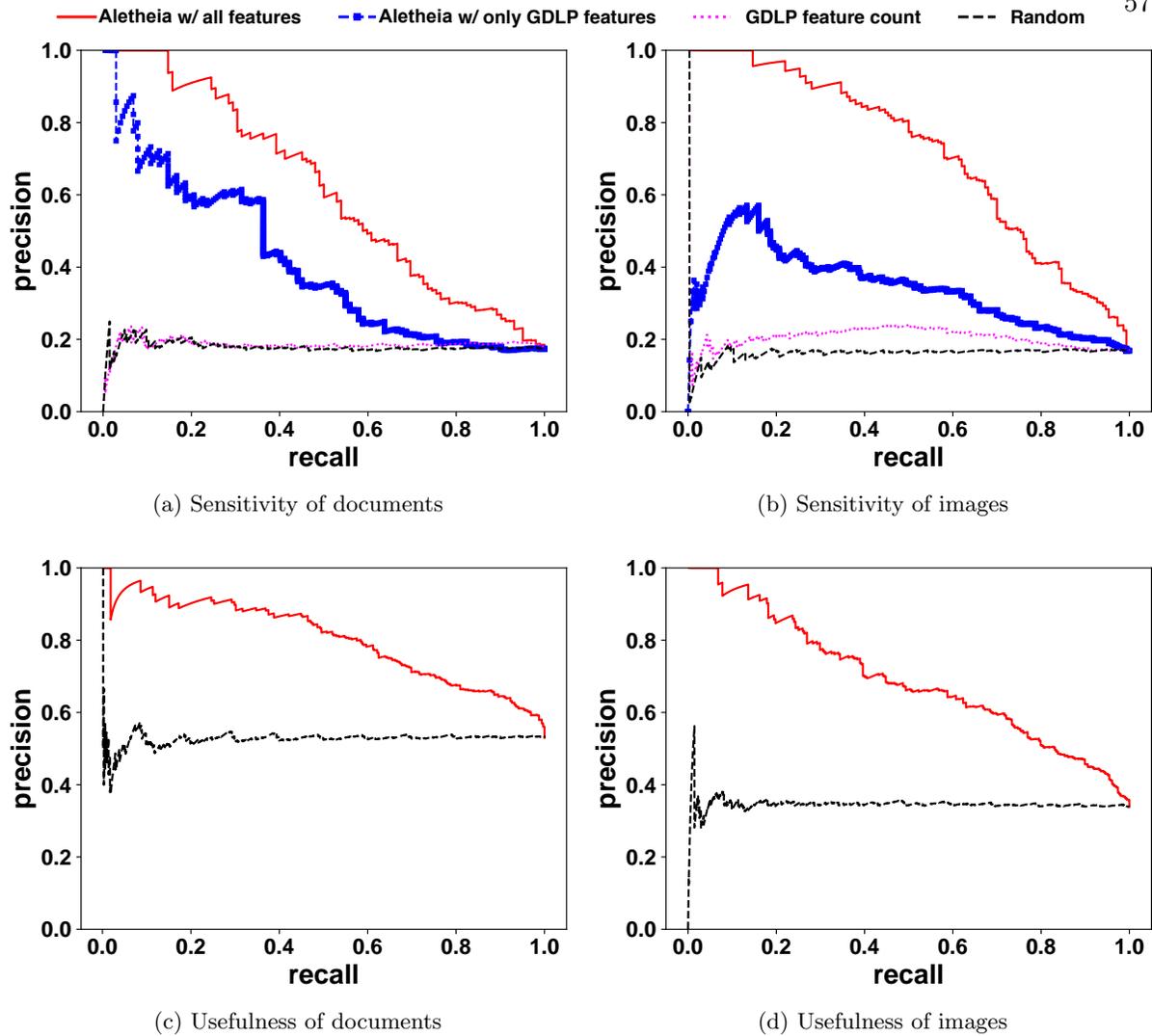


Figure 14: Precision vs. recall for predicting sensitivity and usefulness.

may be indicative of sensitivity for images. We explore user responses and perceptions of sensitivity for in Section 9.6.

**Task 2, Usefulness:** We studied whether it is possible to predict if a user will perceive a file as being not useful for an image or document. Figures 14c–14d show the PRC for predicting *not useful* for images and documents in the usefulness dataset. Here, *Aletheia* performs the best in both tasks, significantly outperforming the baseline classifier.

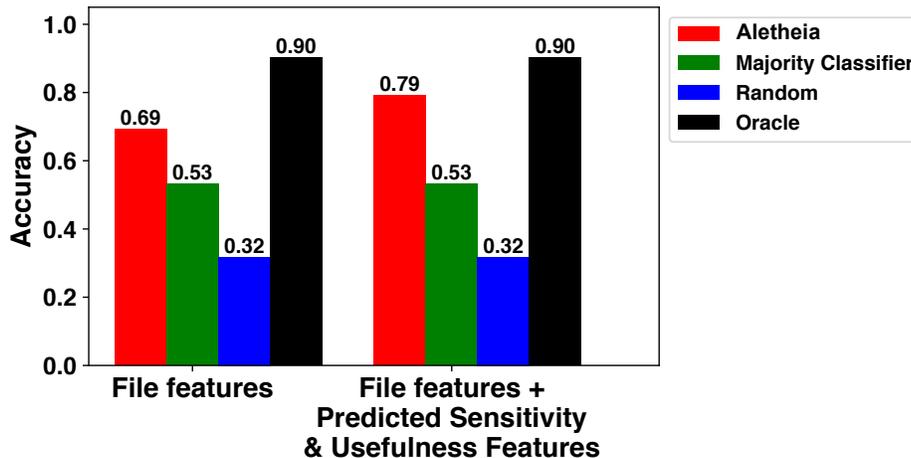


Figure 15: Comparison between directly predicting file management decision and first predicting sensitivity and usefulness for all files.

For documents, the baseline classifier performs reasonably well in this setting, since the distribution of useful documents in the dataset was about even for not useful and useful file labels. *Aletheia* achieves higher AUC value of 0.80 compared to the baseline AUC value of 0.53, an improvement of 51%.

For images, *Aletheia* is more accurate in predicting than the baseline classifier, with an AUC of 0.69 compared to an AUC of 0.35, and improvement of 49%. This may be due to user perception of usefulness, which is less likely to be captured from image features. Our analysis reveals that predicting sensitive images is easier than predicting sensitive documents, while predicting not useful documents is easier than predicting not useful images.

**Task 3, Management Decision:** Next, we study whether we can predict the file management decisions of users. Figure 15 shows the overall accuracy on predicting file management decisions on all files in the testing set. *Aletheia* is compared to a majority classifier and an Oracle that knows users’ responses about a file’s sensitivity and usefulness. We show results for both file management tasks described earlier: predicting only using file features, and the two-step classification setup where we include predicted sensitivity and usefulness as features.

The Oracle shows that if by having the actual responses of perceived usefulness and sensitivity, we can achieve 90% accuracy for predicting file management decisions. Comparing using only file features and predicted sensitivity and usefulness features, we see a significant boost of 10% increase in accuracy. This shows that even

<b>Decision</b>	<b>Correct Predictions</b>	<b>Incorrect Predictions</b>
Keep	91%	9% (delete)
Delete	75%	25% (keep)
Protect	37%	56% (keep), 7% (delete)

Table XI: Accuracy per file management decision and the incorrect (predictions).

without user responses of sensitivity and usefulness of a file, we can use predictions of those perceptions to boost file management prediction accuracy.

Table X shows the breakdown in accuracy across different file management decisions. We are most accurate on keep as-is decisions, since that is the majority class. For cases we mispredict keep as-is decisions, *Aletheia* instead predicts them as *delete*. On the other hand, for delete decisions, *Aletheia* gets a 75% accuracy, while mispredicting delete decisions as *keep*. *Aletheia* does not perform very well on protect labels, with a 37% accuracy. Interestingly, the majority of mispredictions for protect are mislabeled as keep as-is. This shows that *Aletheia* considers protect decisions as closer to keep as-is decisions than delete. The number of protect labels is significantly smaller than the other two labels, which makes it harder to predict.

## 9.6 Understanding Prediction Results

Here, we examine which features were important for each prediction task. Table XI shows the top features identified by each classifier for the sensitivity and usefulness tasks, in order of importance. Word2vec had high feature importance in the classifiers, but since word2vec features are not easy to interpret (94), we do not show them on the list. For documents in the sensitivity task, we noticed that some keywords from the document are important such as military, marriage, and banking keywords. Additionally, personal information such as email, phone number, and date of birth were also important predictors of sensitivity. For images, we observed similar keywords, but also some additional image-only features, such as whether the content was adult or racy.

Task		Features
Sensitivity	Documents	military & marriage <i>keywords</i> , email, banking <i>keywords</i> , sensitive file name, address, phone, date of birth
	Images	access type, racy, violent, spoof, adult finance <i>keywords</i> , medical <i>keywords</i>
Usefulness	Documents	access type, last modifying user, finance <i>keywords</i> report & journal <i>keywords</i> ,
	Images	file size, finance <i>keywords</i> , access type last modifying user, medical <i>keywords</i>
File Management	All Files	usefulness, sensitivity, spoof, account size used space, finance <i>keywords</i> , medical <i>keywords</i>

Table XII: Top features for prediction tasks. Italicized *keywords* are top terms identified via the bag of words collections.

For the usefulness dataset, we saw some similar top features as in the sensitivity dataset, such as access type, and financial keywords. For documents, “report” and “journal” keywords seem to be important in predicting perceived usefulness. For images, medical content was also predictive of usefulness.

Besides the top features in Table XI, word2vec embeddings were also identified as important features, which suggests that text content is central to these prediction tasks. For documents, there was only one word2vec mbedding that represents the entire document, which contributes to classification. However, for images, we considered additional one-hot encoding and word2vec features based on the automatically-identified image labels. Out of those, only word2vec features were identified as important, probably because one-hot encoding of “best guess labels” was too sparse.

Table XI also shows the top most important features for predicting file management decisions in terms of keeping, protecting, and deleting a file. The top two features for predicting this decision were the predicted labels for file usefulness and sensitivity. Using XGBoost, we found that the feature importances of usefulness and sensitivity are 0.40 and 0.11, respectively. This confirms our observation that these two constructs play an important role in file management decisions. Sensitive information in the file, such as spoof content, financial

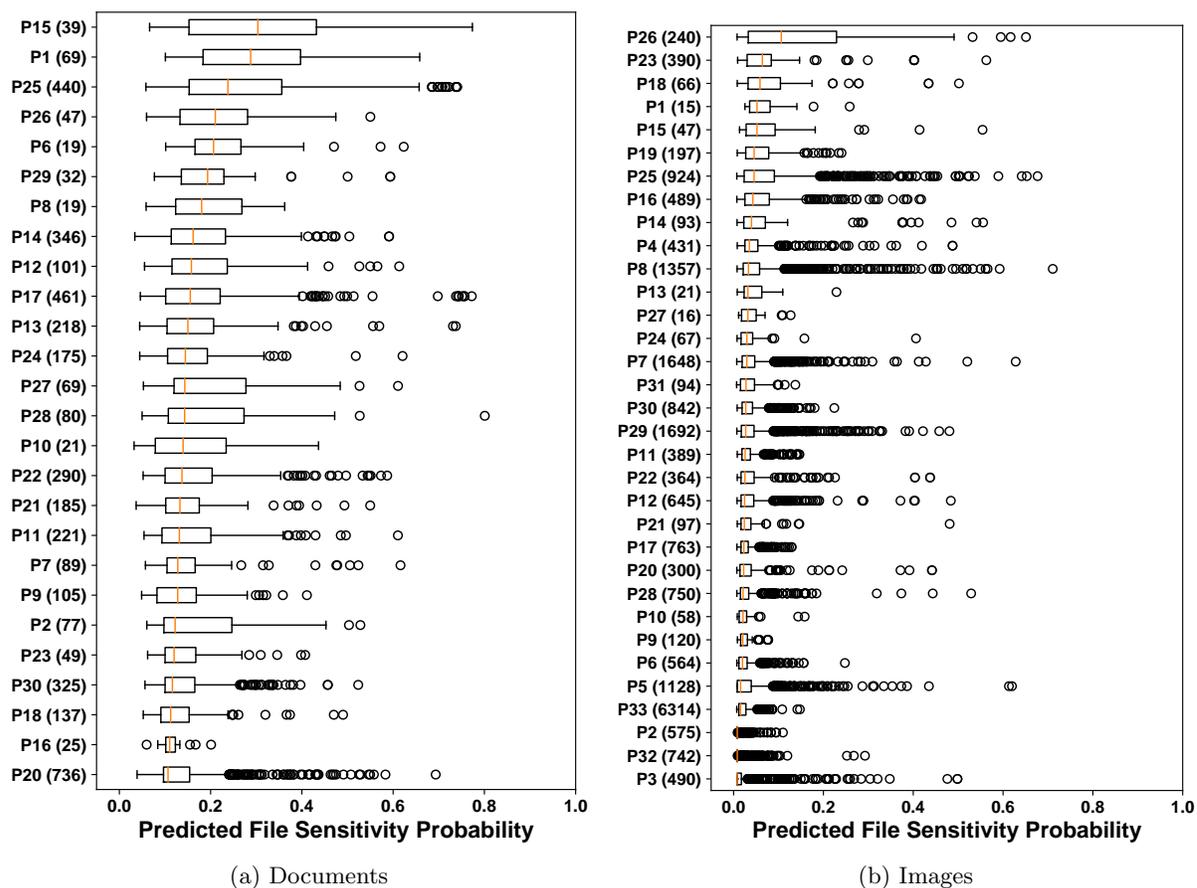


Figure 16: Predicted sensitivity probability for each document and image for every participant.

keywords, and medical content were also important, as well as account features such as account size and how much space the account used.

In order to understand the distribution of sensitive files in each user account in Round 2 data, Figure 16 shows box plots for predicted file sensitivity probability for all documents and images of each user, based on the preliminary classifier predictions. We omitted participants with fewer than 10 documents or images in their accounts. On average, the preliminary classifier predicts the majority of files as having a low probability of being sensitive. Only a small subset of files with a high probability of being sensitive are selected for each user. This means that for many users, we were selecting only a small number of files that the preliminary classifier deemed

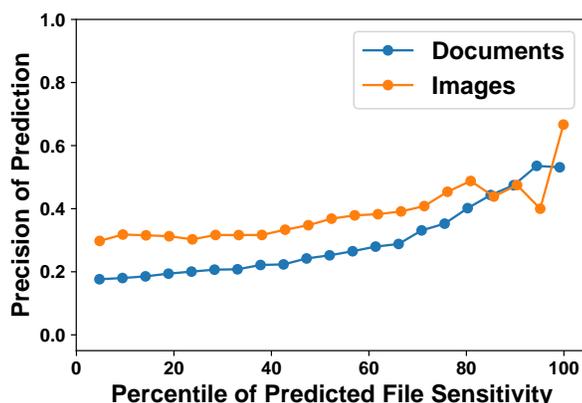


Figure 17: Preliminary classifier prediction precision as a function of predicted file sensitivity.

sensitive with high probability. This results in a low percentage of potentially sensitive files in our final Round 2 dataset.

To explore how accurate our preliminary classifier is at picking sensitive files, we look at the relationship between high file sensitivity probability and the precision of prediction. We ranked the selected files in order of predicted file sensitivity probability and classified files based on a sliding threshold, where everything above the threshold is classified as sensitive. Finally, we compute the precision for predicted files above the threshold on ground truth file sensitivity and report on it in Figure 17 for both images and documents. A higher percentile means a higher threshold for predicted probability of sensitivity. When the threshold for predicted probability is low, we have lower precision (around 30%). With higher predicted probability, our preliminary classifier has better precision of prediction. This shows that the preliminary classifier produces meaningful sensitivity predictions. The increasing precision at higher predicted sensitivity scores indicates that predictions of high sensitivity are more accurate.

We also performed a qualitative evaluation of the final classifier based on the files that were identified as false positives and reason why participants did not consider them sensitive while our classifier ranked them as such. Specifically, we looked at false positives in the top 5 ranked documents and images.

Within documents, 6% of such files contained PII that was obsolete. One participant said *“It’s just a cover letter I had written several years ago and doesn’t contain any good info because the address and phone aren’t*

*good anymore*". From Table XI, we understand that phone numbers and addresses, both are of significance in predicting sensitivity, however, to accurately classify such files, more temporal context of the information is required. Similarly, 3% of the files contained sensitive information but didn't belong to the participants, so they did not state the document as sensitive. Regarding an ex-partner's resume, a participant mentioned "*It might be slightly sensitive to my ex, but not really.*" For a majority of the other documents (70%), participants' responses did not signify a strong element of sensitivity. They mentioned reasons such as information that they did not feel that could compromise them in any manner, or details that were already publicly available.

For images, we noted that most that were classified as sensitive were in fact pictures with faces, memes, or some form of artwork or original content. However, participants do not perceive them as sensitive. For a family photo, a participant said "*This does not reveal any personal information about me, or the person in the photo*". In another example of original artwork, our participant mentioned "*there is nothing sensitive in the file, but I would not want someone stealing the image to use as their own.*" For pictures containing adult content that did not directly affect the participants were not considered sensitive. Regarding a nude photo, a participant mentioned it was not compromising as they were not the subject in the photo.

This investigation reveals that while these selected files were in alignment with our broader definition of sensitivity classification, but in the end, participants made a decision tailored to their personal opinion. Better understanding this requires both additional data collection, and also developing personalized classifiers which account for such personalization. We realize this as a limitation of our current study, and shed light on possible future work in the next Chapter.

## CHAPTER 10

### CONCLUSION

In this chapter, we conclude the findings of the two studies performed. We started this work to understand the broad scope of management of files by users of cloud storage. We specifically looked at if users desired such a form of management, and if so, what are some of the effective mechanisms and metrics we can develop to enable this, especially when the scale of their accounts spans many years of data.

To explore this, we conducted an exploratory study. Results show that participants had many files in the cloud that they had forgotten were there. On showing files, participants wanted to manage at least a quarter of their files by either deleting them or encrypting them. Another contribution of the exploratory study was to develop regression models to see if specific file metrics were significant predictors of management. While the regression models did not provide substantial objective insights, the overall study highlighted the challenges of an automated solution and the need for understanding of the context of management, as well as additional data collection.

To develop more context of file management, we explored participants' free-text responses that explained how their decisions might generalize. Particularly participants wanted to manage files based on their underlying sensitivity and usefulness attributes. This important insight motivated the design of the follow-up study.

In the second study, we focused on exploring the concepts of sensitivity and usefulness and how we could tailor them to perform automated file management. We executed this through a combination of qualitative interviews, data collection surveys, and the development of a machine learning classifier (*Aletheia*). We primarily used the sensitivity/usefulness model from Figure 9 to develop the study. Our results determined that decisions about file management are predicated on several factors, some internal to the user, and some based on the contents of the file. The design of *Aletheia* focused not on directly predicting that decision but rather on predicting perceptions regarding these files that can be inferred using passively collected file metadata, which can then, in turn, be useful in predicting the ultimate file management decision.

Our findings were particularly encouraging for the usefulness part of this model, as using automated inference techniques to first build an understanding of participants' conceptualization of usefulness significantly improved our ability to predict their file management decision. The predicted usefulness was the single most predictive feature for the file management decision classification. This holistic, human-centered approach to create an automated inference highlights the importance of deep qualitative engagement with users during the design of such classifiers.

Not only does this human-centered understanding improve the performance of automated inference, but this approach can also be used to develop a deeper understanding of perceived usefulness and sensitivity for files. Perceptions of usefulness are strongly correlated with future access, while perceptions of sensitivity correlate with the existence of PII, financial information, intimate content, and sentimental value.

While our results suggested a strong correlation between usefulness and desire to delete a given file as well as keeping non-sensitive useful files as-is, there were two notable observations. Firstly, participants' preferences for how to manage useful, sensitive files did not map onto our hypothesized model: decisions to protect useful files were nearly evenly split between sensitive and not sensitive files. Secondly, while not useful files were nearly always deleted, participants still wanted to retain a nontrivial minority of files deemed not useful. This phenomenon suggests that using the concept of usefulness is very helpful for determining whether to retain a given file. Based on this insight, we propose that automated systems should not use such a prediction to make file retention decisions on behalf of the user, but rather should seek confirmation before applying automated management.

While predicting usefulness was incredibly helpful for improving decisions about file management, predicting sensitivity was both less successful and less helpful for supporting the file management inference task. Beyond being harder to accurately predict because the base rate of sensitive files is very low (13%), these phenomena suggest that the relationship between sensitivity and file management is more complex than our hypothesized model. While we were unable to find a performant global (i.e user-agnostic) approach to inferring file sensitivity, future work can explore the possibility that classifiers tuned to individual users' preferences would be able to improve performance on this task.

Within the sensitivity prediction task, our classifier performed better for images than for documents. While this can be an artifact of the underlying data, we hypothesize that some of the significant features for images such as adult, racy, and violent, are evidently easier to automatically detect among images of different users. On the other hand, while for documents, there were standardized classes of significant and clearly identifiable features (e.g., PII and financial information), qualitative responses suggest the presence of a strong temporal aspect associated with such items that our classifier did not take into account. For instance, a resume containing older PII such a previous address would no longer retain its original sensitivity value. This example highlights the essentiality of taking into account temporal sensitivity when predicting management. A similar phenomenon, while it was less pronounced, was also observed in images, where participants described sensitive pictures as having sentimental value (e.g., pictures of children, loved ones). Directly predicting the sensitivity from collected features is unlikely to be successful, and this task certainly merits a more in-depth investigation. Not only is this a difficult task in of itself, but the perception of sentimental value is also unlikely to be user-agnostic, and certainly necessitates the need for classifiers tuned to individual users.

To summarize, in this work, we provide a comprehensive evaluation of the state of users' expectations of cloud storage management as they grow with data, year after year. Using a baseline of perceived file sensitivity and usefulness, we identify file specific characteristics and use them to develop a classification framework to predict management decisions for candidate files in cloud storage accounts. We envision future work that to further improve our understanding of file sensitivity and file management should focus on longitudinal studies to passively observe participants' actions over time rather than actively asking them to make management decisions, and attempt to build a sensitivity/usefulness persona that can account for the variation in perceptions between users. Additionally, our effort was purely focused on an in-lab study rather than an effective interface for cloud archive management; creating an interface that can surface these suggestions efficiently would both allow for operationalizing the results of this project and provide an opportunity for collecting feedback to further improve *Aletheia's* performance.

## CITED LITERATURE

1. Khan, M. T., Hyun, M., Kanich, C., and Ur, B.: Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2018.
2. Khan, M. T., Tran, C., Singh, S., Brackenbury, W., Vasilkov, D., Kanich, C., Ur, B., and Zheleva, E.: Alethia: Helping users automatically find and manage sensitive, expendable files in cloud storages.
3. List of top cloud storage providers 2020. <https://www.trustradius.com/cloud-storage>.
4. Cameron, D.: Apple knew of iCloud security hole 6 months before Celebgate. <http://www.dailydot.com/technology/apple-icloud-brute-force-attack-march/>, September 24 2014.
5. Turner, K.: Hacked dropbox login data of 68 million users is now for sale on the dark web. <https://www.washingtonpost.com/news/the-switch/wp/2016/09/07/hacked-dropbox-data-of-68-million-users-is-now-or-sale-on-the-dark-web/>, Sep 2016.
6. Zetter, K.: Hackers finally post stolen ashley madison data. <https://www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/>, Jun 2017.
7. Phone theft in america: Breaking down the phone theft pandemic. <https://transition.fcc.gov/cgb/events/Lookout-phone-theft-in-america.pdf>.
8. Bergman, O., Whittaker, S., and Frishman, Y.: Let’s get personal: the little nudge that improves document retrieval in the cloud. Journal of documentation, 2019.
9. Mondal, M., Yilmaz, G. S., Hirsch, N., Khan, M. T., Tang, M., Tran, C., Kanich, C., Ur, B., and Zheleva, E.: Moving beyond set-it-and-forget-it privacy settings on social media. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pages 991–1008, 2019.
10. Snyder, P. and Kanich, C.: Cloudsweeper: Enabling data-centric document management for secure cloud archives. In Proceedings of the ACM Cloud Computing Security Workshop, 2013.
11. Google: Cloud data loss prevention. <https://cloud.google.com/dlp/>, 2020.
12. Mell, P., Grance, T., et al.: The nist definition of cloud computing. 2011.
13. Drago, I., Mellia, M., Munafò, M. M., Sperotto, A., Sadre, R., and Pras, A.: Inside Dropbox: Understanding personal cloud storage services. In Proc. IMC, 2012.
14. Kim, B. H., Huang, W., and Lie, D.: Unity: Secure and durable personal cloud storage. In Proc. CCSW, 2012.

15. Cloud Storage Market Size, Share and Growth | Forecast - 2027. <https://www.alliedmarketresearch.com/cloud-storage-market>.
16. Gonçalves, G., Drago, I., Da Silva, A. P. C., Vieira, A. B., and Almeida, J. M.: Modeling the Dropbox client behavior. In Proc. ICC, 2014.
17. Hu, W., Yang, T., and Matthews, J. N.: The good, the bad and the ugly of consumer cloud storage. ACM SIGOPS Operating Systems Review, 44(3):110–115, 2010.
18. Graham Cluley: Dropbox users leak tax returns, mortgage applications and more. <https://www.grahamcluley.com/dropbox-box-leak/>, Accessed 2017.
19. Johnson, E.: Lost in the cloud: Cloud storage, privacy, and suggestions for protecting users' data. Stan. L. Rev., 69:867, 2017.
20. Clark, J. W., Snyder, P., McCoy, D., and Kanich, C.: I Saw Images I Didn't Even Know I Had: Understanding User Perceptions of Cloud Storage Privacy. In Proc. CHI, 2015.
21. Stark, L. and Tierney, M.: Lockbox: Mobility, privacy and values in cloud storage. Ethics and Information Technology, 16(1):1–13, 2014.
22. Ion, I., Sachdeva, N., Kumaraguru, P., and Čapkun, S.: Home is safer than the cloud!: Privacy concerns for consumer cloud storage. In Proc. SOUPS, 2011.
23. Sultan, N. A.: Reaching for the cloud: How smes can manage. International journal of information management, 31(3):272–278, 2011.
24. Chow, R., Golle, P., Jakobsson, M., Shi, E., Staddon, J., Masuoka, R., and Molina, J.: Controlling data in the cloud: Outsourcing computation without outsourcing control. In Proc. CCSW, 2009.
25. Schindler, E.: Cloud development survey. Evans Data Corporation Strategic Reports, July 2010.
26. Mijuskovic, A. and Ferati, M.: User awareness of existing privacy and security risks when storing data in the cloud. In Proc. e-Learning, 2015.
27. Arpaci, I., Kilicer, K., and Bardakci, S.: Effects of security and privacy concerns on educational use of cloud services. Computers in Human Behavior, 45:93–98, 2015.
28. Ohlhausen, M. K.: Painting the privacy landscape: Informational injury in ftc privacy and data security cases. [https://www.ftc.gov/system/files/documents/public\\_statements/1255113/privacy\\_speech\\_mkohlhausen.pdf](https://www.ftc.gov/system/files/documents/public_statements/1255113/privacy_speech_mkohlhausen.pdf), 2017.
29. Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., and Dell, N.: “a stalker’s paradise”: How intimate partner abusers exploit technology. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, page 667. ACM, 2018.
30. Mitchell, K. J., Jones, L. M., Finkelhor, D., and Wolak, J.: Trends in unwanted online experiences and sexting. 2014.

31. Livingstone, S., Haddon, L., Görzig, A., and Ólafsson, K.: Risks and safety on the internet. The perspective of European children. Full findings and policy implications from the EU Kids Online survey of, pages 9–16, 2011.
32. Kokolakis, S.: Privacy attitudes and privacy behaviour. Comput. Secur., 64(C):122–134, January 2017.
33. Aguirre, X., Woodward, B., and Martin, N.: Perception of privacy through generations. Issues in Information Systems, 19(1), 2018.
34. Miyazaki, A. D. and Fernandez, A.: Consumer perceptions of privacy and security risks for online shopping. Journal of Consumer affairs, 35(1):27–44, 2001.
35. Douglas, D. M.: Doxing: a conceptual analysis. Ethics and information technology, 18(3):199–210, 2016.
36. Snyder, P., Doerfler, P., Kanich, C., and McCoy, D.: Fifteen minutes of unwanted fame: detecting and characterizing doxing. pages 432–444, 11 2017.
37. Eterovic-Soric, B., Choo, K.-K. R., Ashman, H., and Mubarak, S.: Stalking the stalkers—detecting and deterring stalking behaviours using technology: A review. computers & security, 70:278–289, 2017.
38. Ellison, L. and Akdeniz, Y.: Cyber-stalking: the regulation of harassment on the internet. Criminal Law Review, 29:29–48, 1998.
39. Brandom, R.: The capital one breach is more complicated than it looks. <https://www.theverge.com/2019/7/31/20748886/capital-one-breach-hack-thompson-security-data>, Jul 2019.
40. John, A. S.: Equifax data breach: What consumers need to know. <https://www.consumerreports.org/privacy/what-consumers-need-to-know-about-the-equifax-data-breach/>.
41. Sleeper, M., Melicher, W., Habib, H., Bauer, L., Cranor, L. F., and Mazurek, M. L.: Sharing personal content online: Exploring channel choice and multi-channel behaviors. In Proc. CHI, 2016.
42. Brackenbury, W.: Files of a feather flock together?
43. Zhao, X., Salehi, N., Naranjit, S., Alwaalan, S., Volda, S., and Cosley, D.: The many faces of facebook: Experiencing social media as performance, exhibition, and personal archive. In Proc. CHI, 2013.
44. Ayalon, O. and Toch, E.: Retrospective privacy: Managing longitudinal privacy in online social networks. In Proc. SOUPS, 2013.
45. Bauer, L., Cranor, L. F., Komanduri, S., Mazurek, M. L., Reiter, M. K., Sleeper, M., and Ur, B.: The post anachronism: The temporal dimension of facebook privacy. In Proc. WPES, 2013.
46. Mondal, M., Messias, J., Ghosh, S., Gummadi, K. P., and Kate, A.: Longitudinal privacy management in social media: The need for better controls. IEEE Internet Computing, 21(3):48–55, 2017.
47. Google: Google privacy and terms. <https://policies.google.com/privacy>, 2020.

48. Facebook: Community standards. <https://www.facebook.com/communitystandards/>, 2019.
49. Twitter: The twitter rules. <https://help.twitter.com/en/rules-and-policies/twitter-rules>, 2019.
50. Peddinti, S. T., Korolova, A., Bursztein, E., and Sampemane, G.: Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In 2014 IEEE Symposium on Security and Privacy, pages 493–508. IEEE, 2014.
51. Vitale, F., Janzen, I., and McGrenere, J.: Hoarding and minimalism: Tendencies in digital data preservation. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2018.
52. Vitale, F., Odom, W., and McGrenere, J.: Keeping and discarding personal data: Exploring a design space. In Proceedings of the 2019 on Designing Interactive Systems Conference, pages 1463–1477. ACM, 2019.
53. Ramokapane, K. M., Rashid, A., and Such, J.: “i feel stupid i can’t delete...”: A study of users’ cloud deletion practices and coping strategies. In Proc. SOUPS, 2017.
54. Murillo, A., Kramm, A., Schnorf, S., and De Luca, A.: “if i press delete, it’s gone”: User understanding of online data deletion and expiration. In Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security, SOUPS ’18, page 329–339, USA, 2018. USENIX Association.
55. Axtell, B. and Munteanu, C.: Back to real pictures: A cross-generational understanding of users’ mental models of photo cloud storage. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(3):74, 2019.
56. Kang, R., Dabbish, L., Fruchter, N., and Kiesler, S.: “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 (SOUPS) 2015, pages 39–52, 2015.
57. Turczyk, L. A., Heckmann, O., and Steinmetz, R.: File valuation in information lifecycle management. In Proceedings of the Thirteenth Americas Conference on Information Systems, Keystone, Colorado, 2007.
58. Wijnhoven, F., Amrit, C., and Dietz, P.: Value-based file retention: File attributes as file value and information waste indicators. Journal of Data and Information Quality (JDIQ), 4(4):1–17, 2014.
59. Lansdale, M. W.: The psychology of personal information management. Applied ergonomics, 19(1):55–66, 1988.
60. Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., and Neuwirth, C.: Innovation in extremis: Evolving an application for the critical work of email and information management. In Proc. DIS, 2002.
61. Barreau, D. K.: Context as a factor in personal information management systems. Journal of the American Society for Information Science, 46(5):327, 1995.

62. Boardman, R. P.: Improving tool support for personal information management. Doctoral dissertation, University of London, 2004.
63. Bergman, O., Boardman, R., Gwizdka, J., and Jones, W.: Personal information management. In Proc. CHI Extended Abstracts, 2004.
64. Dumais, S., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R., and Robbins, D. C.: Stuff i've seen: A system for personal information retrieval and re-use. In ACM SIGIR Forum, volume 49, pages 28–35, 2016.
65. Kaptelinin, V.: Umea: Translating interaction histories into project contexts. In Proc. CHI, 2003.
66. Balter, O.: Strategies for organising email messages. In Proc. HCI, 1997.
67. Whittaker, S. and Sidner, C.: Email overload: Exploring personal information management of email. In Proc. CHI, 1996.
68. Whittaker, S., Bellotti, V., and Gwizdka, J.: Email in personal information management. Communications of the ACM, 49(1):68–73, 2006.
69. Ayodele, T., Akmayeva, G., and Shoniregun, C. A.: Machine learning approach towards email management. In Proc. World Congress on Internet Security (WorldCIS), pages 106–109, 2012.
70. Barreau, D. and Nardi, B. A.: Finding and reminding: File organization from the desktop. ACM SigChi Bulletin, 27(3):39–43, 1995.
71. Fang, L. and LeFevre, K.: Privacy wizards for social networking sites. In WWW, 2010.
72. Ghazinour, K., Matwin, S., and Sokolova, M.: Monitoring and recommending privacy settings in social networks. In EDBT Workshops, 2013.
73. Liu, K. and Terzi, E.: A Framework for Computing the Privacy Scores of Users in Online Social Networks. TKDD, 5(1):6, 2010.
74. Zheleva, E., Terzi, E., and Getoor, L.: Privacy in Social Networks. Synthesis Lectures on Data Mining and Knowledge Discovery, 3(1):1–85, 2012.
75. Zheleva, E. and Getoor, L.: To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In Proc. WWW, 2009.
76. Lindamood, J., Heatherly, R., Kantarcioglu, M., and Thuraisingham, B.: Inferring Private Information Using Social Network Data. In Proc. WWW, 2009.
77. Kosinski, M., Stillwell, D., and Graepel, T.: Private Traits and Attributes are Predictable From Digital Records of Human Behavior. PNAS, 110(15):5802–5805, 2013.
78. Gong, N. Z. and Liu, B.: You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In Proc. USENIX Security, 2016.

79. Garcia, D.: Leaking Privacy and Shadow Profiles in Online Social Networks. Science Advances, 3(8), 2017.
80. Stutzman, F., Gross, R., and Acquisti, A.: Silent listeners: The evolution of privacy and disclosure on facebook. Journal of Privacy and Confidentiality, 4(2):7–41, 2013.
81. Fiesler, C., Dye, M., Feuston, J. L., Hiruncharoenvate, C., Hutto, C. J., Morrison, S., Roshan, P. K., Pavalanathan, U., Bruckman, A. S., De Choudhury, M., and Gilbert, E.: What (or who) is public?: Privacy settings and social media content sharing. In CSCW, pages 567–580, 2017.
82. Liu, B., Andersen, M., Schaub, F., Almuhimedi, H., Zhang, S., Sadeh, N., Acquisti, A., and Agarwal, Y.: Follow my recommendations: A personalized privacy assistant for mobile app permissions. In SOUPS, 2016.
83. Tonge, A. and Caragea, C.: Dynamic deep multi-modal fusion for image privacy prediction. In The World Wide Web Conference, pages 1829–1840. ACM, 2019.
84. Dropbox paper. <https://www.dropbox.com/paper>.
85. Oauth 2.0. <https://oauth.net/2/>.
86. Kollmar, F.: Cloud storage report 2017. <https://blog.cloudrail.com/cloud-storage-report-2017/>, 2017.
87. Houston, D. and Ferdowsi, A.: Celebrating half a billion users. <https://blogs.dropbox.com/dropbox/2016/03/500-million/>, 2016.
88. Dupree, J. L., Devries, R., Berry, D. M., and Lank, E.: Privacy personas: Clustering users via attitudes and behaviors toward security practices. In Proc. CHI, 2016.
89. Google: Cloud vision. <https://cloud.google.com/vision/>, 2020.
90. Google: Speech to text. <https://cloud.google.com/speech-to-text>, 2020.
91. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
92. Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794. ACM, 2016.
93. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.
94. Levy, O. and Goldberg, Y.: Dependency-based word embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 302–308, 2014.

## APPENDIX

### Exploratory Study Survey Instrument

Questions prefixed with the  $\Rightarrow$  symbol indicate conditional-branched questions.

#### 1. Generic Questions

**G-1** For approximately how long have you had the *Cloud Storage* account you are using for this study?

- Less than 1 year
- At least 1 year, but less than 2 years
- At least 2 years, but less than 3 years
- At least 3 years, but less than 4 years
- At least 4 years, but less than 5 years
- More than 5 years

**G-11** Cloud storage providers offer both free accounts and paid accounts, where the latter offers more storage space. Do you use a free *Cloud Storage* account or a paid *Cloud Storage* account?

- Free account
- Paid account
- I'm not sure

$\Rightarrow$  **G-12** How much do you pay per month?

**G2** How often do you use this *Cloud Storage* account for work or school purposes?

- At least once a week
- At least once a month, but less than once a week
- At least once a year, but less than once a month
- Less than once a year, but sometimes
- I do not use it for work or school purposes

**G-3** How often do you use this *Cloud Storage* account for personal purposes (i.e., for purposes other than for work or school)?

- At least once a week
- At least once a month, but less than once a week
- At least once a year, but less than once a month
- Less than once a year, but sometimes
- I do not use it for personal purposes

**G-4 I use this *Cloud Storage* account for the following purposes: (Check all that apply)**

- Collaborating with co-workers or classmates by jointly creating and editing files
- Collaborating with friends and family by jointly creating and editing files
- Sharing files that I have created with co-workers, classmates, or other professional contacts
- Sharing files that I have created with family and friends
- Backing up files related to my job, school, or career
- Backing up files that are not related to my job, school, or career
- Other

**G-5 There are multiple ways you can access files in your *Cloud Storage* account. One of these ways is by installing *Cloud Storage* software on your computer so that certain folders are automatically synced with your *Cloud Storage* account. How often do you access (view or edit) files or folders on your computer that are automatically synced with your *Cloud Storage* account?**

- Daily or more frequently
- Every few days
- Weekly
- Monthly
- Less than once a month, but sometimes
- Never

**G-6 Another way to access files in your *Cloud Storage* account is by using a web browser like Chrome, Firefox, or Safari to log into the *Cloud Storage* website. How often do you log into the *Cloud Storage* website using this account?**

- Daily or more frequently
- Every few days
- Weekly
- Monthly
- Less than once a month, but sometimes
- Never

**G-7 Yet another way to access files in your *Cloud Storage* account is by using an app on your smartphone (iPhone or Android). How often do you use a smartphone app to access files or folders stored in this *Cloud Storage* account?**

- Daily or more frequently
- Every few days
- Weekly
- Monthly
- Less than once a month, but sometimes
- Never, though I do use a smartphone
- Never; I do not use a smartphone

**G-8** The following two questions concern the following distinction: A file stored locally is accessible on your computer (i.e., stored on the hard drive) even if you are not connected to the Internet. A file stored in the cloud is accessible only if you are connected to the Internet. Note that a given file might be stored both locally and in the cloud.

**G8-1** Which statement describes your current situation for cloud files?

- All of my cloud files are also stored locally on my computer
- Most of my cloud files are also stored locally on my computer
- Some of my cloud files are also stored locally on my computer
- None of my cloud files are also stored locally on my computer

**G8-2** Which statement best describes your current situation for files stored locally?

- All of my locally stored files are also accessible in the cloud via *Cloud Storage*
- Most of my locally stored files are also accessible in the cloud via *Cloud Storage*
- Some of my locally stored files are also accessible in the cloud via *Cloud Storage*
- None of my locally stored files are also accessible in the cloud via *Cloud Storage*

**G-9** On average, how often do you run out of space on your *Cloud Storage* account?

- I am almost always out of storage space
- At least once a month
- At least once a year, but less than once a month
- Less than once a year, but sometimes
- I have never run out of storage space
- I don't know

**G-10** On average, how often do you organize your *Cloud Storage* by deleting unnecessary files, moving files to different folders, or performing similar clean-up tasks?

- At least once a week
- At least once a month, but less than once a week
- At least once a year, but less than once a month
- Less than once a year, but sometimes
- I have never organized my *Cloud Storage*
- I don't know

**G-13** Overall, which of the following cloud services do you use? (Check all that apply)

- Amazon Cloud Drive
- Apple iCloud
- Box
- Dropbox
- Google Drive
- Microsoft OneDrive
- SpiderOak One
- Other

## 2. File-Specific Questions

**CF-1** After looking at this file, do you know what it is? (Note: You might not know what it is if the file was automatically created or automatically saved to your cloud storage.)

- Yes
- No

⇒ **CF-11** Prior to this survey, I remembered that this file was stored on any device or service I use.

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly disagree

⇒ **CF-12** Prior to this survey, I remembered that this file was stored in my *Cloud Storage*.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

⇒ **CF-13** As far as you can remember, why did you originally store this file on *Cloud Storage*?

⇒ **CF-14** As far as you can remember, when did you originally store this file on *Cloud Storage*?

- Within the last week
- At least a week ago, but less than a month ago
- At least a month ago, but less than a year ago
- At least a year ago, but less than five years ago
- At least five years ago
- I don't know
- As far as I remember, I did not store it on *Cloud Storage*

⇒ **CF-15** As far as you remember, when is the last time you accessed (viewed or modified) this file?

- Less than a week ago
- Over 1 week ago, but less than 1 month ago
- Over 1 month ago, but less than 1 year ago
- Over 1 year ago, but less than 5 years ago
- Over 5 years ago
- As far as I know, I have never accessed this file
- I don't remember

⇒ **CF-16** When do you next expect to access (view or modify) this file in the future?

- Within the next week
- Over 1 week from now, but less than 1 month from now
- Over 1 month from now, but less than 1 year from now
- Over 1 year from now, but less than 5 years from now
- Over 5 years from now, but eventually
- Never

**CF-2** Which of these statements best characterizes what you would like to happen to this file?

- I would like to keep this file stored as-is in my *Cloud Storage*.
- I would like to keep only an encrypted version of this file in my *Cloud Storage*.
- I would like to delete this file from my *Cloud Storage*.

⇒ **CF-21** Files could potentially be stored in the cloud in a way that saves energy. However, this would mean that the file could only be accessed with some delay, rather than instantaneously. When I next try to access this file...

- ...no delay in the file being available is acceptable
- ...a delay of up to a few minutes in being able to access the file is acceptable
- ...a delay of up to a few hours in being able to access the file is acceptable
- ...a delay of up to a few days in being able to access the file is acceptable

⇒ **CFE-1** It is important to me that this file is encrypted, rather than remaining as-is in my *Cloud Storage*.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

⇒ **CFD-1** It is important to me that this file is deleted, rather than remaining as-is in my *Cloud Storage*.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

⇒ **CFA-1** Why would you want to continue storing this file as-is on *Cloud Storage*?

⇒ **CFE-2** Why would you want to keep an encrypted version of this file on *Cloud Storage*?

⇒ **CFD-2** Why would you want to delete this file from *Cloud Storage*?

**CF-3** It is important to me to keep this file safe from unauthorized access.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

**CF-4** It is important to me that I never lose the ability to access this file.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

**CF-5** As far as you know, do you have a copy of this file on any other device or service you use?

- Yes, I have another copy of the file somewhere
- No, I do not have any other copies of this file
- I'm not sure

⇒ **CFD-3** Which of the following two statements better describes what you would want to happen?

- Although I would like to delete this file from my *Cloud Storage* account, I would want to keep a copy of the file on a local device (e.g., my computer or smartphone)
- I would like to delete this file from my *Cloud Storage* account, and I would not want to keep a copy of the file on any of my local devices

**CF-6** Are there any other files stored in your *Cloud Storage* account for which you would want to apply the same file-management decision (keep as-is, encrypt, delete) as for this file?

- Yes
- No
- I'm not sure

⇒ **CF-61** For what other files in your *Cloud Storage* would you want to apply the same file-management decision? Please describe those files using whatever language you use to think about them, rather than constraining yourself to the current *Cloud Storage* interface.

⇒ **CF-62** Why would you not want to apply the same file-management decision from this file to other files in your *Cloud Storage*?

### Questions For Files Shared with Individuals

**CSP-1** For each of these people with whom the file is shared, indicate below whether you know who the person is.

	I know who this is, and I have talked to them within the last year	I know who this is, but I have not talked to them in over a year	I do not know who this is
Member A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Member B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Member C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**CSP-2** For each of these people, indicate below whether you would want to keep sharing this particular file with that person, stop sharing this particular file with that person, or whether it doesn't matter to you.

	Definitely keep sharing	Doesn't matter	Definitely stop sharing
Member A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Member B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Member C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**CSP-3** To your knowledge, were you the person who originally shared this file with those people?

- I am the person who shared this file with all of those people
- I am the person who shared this file with some, but not all, of those people
- I am not the person who shared this file with any of those people
- I don't know

⇒ **CSP-11** You indicated that you want to keep sharing this file with at least one other person. Why do you want to keep sharing this file with them?

⇒ **CSP-12** You indicated that you want to stop sharing this file with at least one other person. Why do you want to stop sharing this file with them?

**CSP-31** If you remember, when did you first share this file with other people?

- Less than a week ago (1)
- Over 1 week ago, but less than 1 month ago (2)
- Over 1 month ago, but less than 1 year ago (3)
- Over 1 year ago, but less than 5 years ago (4)
- Over 5 years ago (5)
- I don't know (6)

⇒ **CSP-32 (Optional)** If you remember, why did you originally share this file with *MamberA*?

⇒ **CSP-33 (Optional)** If you remember, why did you originally share this file with *MamberB*?

⇒ **CSP-34 (Optional)** If you remember, why did you originally share this file with *MamberC*?

**CSP-4** If anyone other than me changes (modifies or deletes) the file, my copy of the file should also reflect their changes.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

**CSP-41** Why?

**CSP-5** If I change (modify or delete) this file, other people's copies of the file should also reflect my changes.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

**CSP-51** Why?

### Questions For Files Shared via Link

**CSI-1 To your knowledge, were you the person who created a shareable link for this file?**

- Yes, I am the person who created the link for sharing this file
- No, I am not the person who created the link for sharing this file
- I don't know

⇒ **CSI-11 To your knowledge, with how many people have you shared the link to access this file?**

- No one other than yourself
- 1 - 5 people
- 6 - 10 people
- 11 - 15 people
- 16 - 20 people
- More than 20 people
- I don't know

**CSI-2 Do you want to keep sharing this particular file with others using a link, stop sharing this particular file with others using a link, or does it not matter to you?**

- Definitely keep sharing using a link
- Doesn't matter
- Definitely stop sharing using a link

⇒ **CSI-21 You indicated that you want to keep sharing this file using a link. Why do you want to keep sharing this file?**

⇒ **CSI-22 You indicated that you want to stop sharing this file using a link. Why do you want to stop sharing this file?**

⇒ **CSI-12 If you remember, when did you set this file to be shared using a link?**

- Less than a week ago
- Over 1 week ago, but less than 1 month ago
- Over 1 month ago, but less than 1 year ago
- Over 1 year ago, but less than 5 years ago
- Over 5 years ago
- I don't know

⇒ **CSI-13 (Optional) If you remember, why did you originally share this file using a link?**

**CSI-3 If anyone other than me changes (modifies or deletes) the file, my copy of the file should also reflect their changes.**

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

**CSI-31 Why?**

**CSI-4 If I change (modify or delete) this file, other people's copies of the file should also reflect my changes.**

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

**CSI-41 Why?**

### 3. Demographic Questions

**DE** (Regardless of whether or not I would want to encrypt any of the files I saw in today's study,) it would be helpful if I could specify that a file on my *Cloud Storage* should be automatically encrypted.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

⇒ **DE-11** Why would it be helpful?

⇒ **DE-12** How, if at all, could *Cloud Storage* identify files or folders in your account that should be automatically encrypted?

⇒ **DE-2** Why would it not be helpful?

**DD** It would be helpful if I could choose that certain files or folders would automatically and permanently delete themselves from my *Cloud Storage* account after a period of time I specify.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

⇒ **DD-11** Why would it be helpful??

⇒ **DD-12** How, if at all, could *Cloud Storage* automatically identify files or folders in your account that should be automatically deleted??

⇒ **DD-2** Why would it not be helpful?

**DA** It would be helpful if I could specify that certain files or folders would automatically move to an archive (saving energy, but causing a delay when I try to access the file) after a period of time I specify.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

⇒ DA-11 Why would it be helpful?

⇒ DA-12 How, if at all, could *Cloud Storage* automatically identify files or folders in your account that should be automatically moved to an energy-saving archive?

⇒ DA-2 Why would it not be helpful?

DC (Optional) Do you have any other comments about anything in today's survey?

DP-1 With what gender do you identify?

- Male
- Female
- Other
- Prefer not to answer

DP-2 Are you majoring in, or do you have a degree or job in, computer science, computer engineering, information technology, or a related field?

- Yes
- No
- Prefer not to answer

DP-3 How old are you?

DP-4 What is your occupation?

## Exploratory Study Regression Tables

The results of a mixed-effects logistic regression to identify what factors were correlated with **recognizing what the file is** (baseline: not recognized). Non-italicized values in the baseline column specify the baseline category for terms representing categorical variables. Italicized values in the baseline column indicate the units for numerical terms. Significant p-values are shown in bold.

Table XIII: Appendix - Factors correlated with file recognition.

Factor	Baseline / <i>Units</i>	Coefficient	Std. Error	z value	p
Service: Dropbox	Google Drive	-0.727	2.219	-0.328	0.743
File Type: Document	Other	1.997	0.488	4.090	<.001
File Type: Image	Other	0.940	0.424	2.215	<b>0.027</b>
File Type: Spreadsheet	Other	1.209	0.896	1.349	0.177
File Type: Video	Other	1.483	1.070	1.386	0.166
Access: Editor	Owner	-2.143	0.653	-3.281	<b>0.001</b>
Access: Viewer	Owner	-1.690	0.664	-2.546	<b>0.011</b>
Days Since Modified	<i>log10(days+1)</i>	-0.308	0.285	-1.082	0.279
File Size	<i>log10(bytes)</i>	0.264	0.142	1.862	0.062
Shared	Not shared	-0.058	0.541	-0.107	0.915
Account Age	<i>Years</i>	0.274	0.166	1.654	0.098
Participant Tech. Background	No	0.350	0.429	0.816	0.414
Participant Age	<i>Years</i>	0.016	0.021	0.772	0.440
Account for Work Purposes	No	0.491	0.427	1.150	0.250
Account for Personal Purposes	No	-0.147	0.528	-0.278	0.781
Service: Dropbox * File Type: Document	Google Drive, Other	0.050	0.866	0.058	0.954
Service: Dropbox * File Type: Image	Google Drive, Other	0.425	0.733	0.580	0.562
Service: Dropbox * File Type: Spreadsheet	Google Drive, Other	0.152	1.499	0.102	0.919
Service: Dropbox * File Type: Video	Google Drive, Other	0.187	1.702	0.110	0.912
Service: Dropbox * Access Type: Editor	Google Drive, Owner	2.983	1.258	2.371	<b>0.018</b>
Service: Dropbox * Access Type: Viewer	Google Drive, Owner	1.038	1.688	0.615	0.539
Service: Dropbox * Days Since Modified	Google Drive, N/A	-0.538	0.591	-0.911	0.362
Service: Dropbox * File Size	Google Drive, N/A	0.276	0.213	1.297	0.195
Service: Dropbox * Shared	Google Drive, Not	0.754	0.883	0.853	0.393

The results of a mixed-effects ordinal regression to identify what factors were correlated with **remembering that the file was stored in the cloud**, which was recorded on a five-point Likert scale coded as an integer from -2 (the participant strongly disagrees that they remembered the file was stored in the cloud) to 2 (the participant strongly agrees that they remembered the file was stored in the cloud). Non-italicized values in the baseline column specify the baseline category for terms representing categorical variables. Italicized values in the baseline column indicate the units for numerical terms. Significant p-values are shown in bold.

Table XIV: Appendix - Factors correlated with file remembrance.

Factor	Baseline / <i>Units</i>	Coefficient	Std. Error	z value	p
Service: Dropbox	Google Drive	-1.444	1.111	-1.300	0.193
File Type: Document	Other	0.102	0.220	0.465	0.642
File Type: Image	Other	-0.150	0.002	-88.197	<.001
File Type: Spreadsheet	Other	-0.063	0.599	-0.104	0.917
File Type: Video	Other	1.467	0.655	2.239	<b>0.025</b>
Access: Editor	Owner	-1.067	0.428	-2.493	<b>0.013</b>
Access: Viewer	Owner	-3.09	0.455	-6.789	<.001
Days Since Modified	<i>log10(days+1)</i>	-0.658	0.002	-385.835	<.001
File Size	<i>log10(bytes)</i>	0.058	0.002	34.251	<.001
Shared	Not shared	0.375	0.002	220.280	<.001
Account Age	<i>Years</i>	-0.126	0.002	-73.893	<.001
Participant Tech. Background	No	-0.459	0.433	-1.061	0.289
Participant Age	<i>Years</i>	0.011	0.002	6.491	<.001
Account for Work Purposes	No	-0.382	0.002	-212.773	<.001
Account for Personal Purposes	No	0.726	0.002	404.017	<.001
Service: Dropbox * File Type: Document	Google Drive, Other	0.025	0.483	0.051	0.959
Service: Dropbox * File Type: Image	Google Drive, Other	0.106	0.403	0.263	0.793
Service: Dropbox * File Type: Spreadsheet	Google Drive, Other	0.682	0.988	0.690	0.490
Service: Dropbox * File Type: Video	Google Drive, Other	-2.348	0.917	-2.560	<b>0.010</b>
Service: Dropbox * Access Type: Editor	Google Drive, Owner	1.549	0.688	2.250	<b>0.024</b>
Service: Dropbox * Access Type: Viewer	Google Drive, Owner	4.084	1.187	3.441	<.001
Service: Dropbox * Days Since Modified	Google Drive, N/A	-0.294	0.259	-1.133	0.257
Service: Dropbox * File Size	Google Drive, N/A	0.294	0.095	3.086	<b>0.002</b>
Service: Dropbox * Shared	Google Drive, Not	-0.493	0.428	-1.154	0.249

The results of a mixed-effects logistic regression to identify what factors were correlated with **expressing a preference to delete the file shown**, as opposed to keeping the file as-is. Files the participant wanted to encrypt are excluded from this model. Non-italicized values in the baseline column specify the baseline category for terms representing categorical variables. Italicized values in the baseline column indicate the units for numerical terms. Significant p-values are shown in bold.

Table XV: Appendix - Factors correlated with preferences for file deletion.

Factor	Baseline / <i>Units</i>	Coefficient	Std. Error	z value	p
Service: Dropbox	Google Drive	-2.342	1.556	-1.505	0.132
File Type: Document	Other	-0.375	0.335	-1.121	0.262
File Type: Image	Other	-0.226	0.337	-0.671	0.502
File Type: Spreadsheet	Other	1.233	0.696	1.772	0.076
File Type: Video	Other	-1.143	0.683	-1.673	0.094
Access: Editor	Owner	1.379	0.518	2.665	<b>0.008</b>
Access: Viewer	Owner	2.054	0.535	3.838	<b>&lt;.001</b>
Days Since Modified	<i>log10(days+1)</i>	0.077	0.189	0.407	0.684
File Size	<i>log10(bytes)</i>	-0.149	0.105	-1.419	0.156
Shared	Not shared	-0.361	0.359	-1.006	0.314
Account Age	<i>Years</i>	-0.186	0.142	-1.316	0.188
Participant Tech. Background	No	-0.129	0.362	-0.355	0.722
Participant Age	<i>Years</i>	-0.018	0.017	-1.012	0.312
Account for Work Purposes	No	-0.045	0.361	-0.123	0.902
Account for Personal Purposes	No	-0.420	0.435	-0.964	0.335
Service: Dropbox * File Type: Document	Google Drive, Other	1.024	0.631	1.623	0.105
Service: Dropbox * File Type: Image	Google Drive, Other	1.208	0.602	2.007	<b>0.045</b>
Service: Dropbox * File Type: Spreadsheet	Google Drive, Other	-0.158	1.278	-0.123	0.902
Service: Dropbox * File Type: Video	Google Drive, Other	1.350	1.073	1.258	0.208
Service: Dropbox * Access Type: Editor	Google Drive, Owner	-2.924	0.864	-3.385	<b>&lt;.001</b>
Service: Dropbox * Access Type: Viewer	Google Drive, Owner	-3.322	1.624	-2.045	<b>0.041</b>
Service: Dropbox * Days Since Modified	Google Drive, N/A	0.351	0.355	0.989	0.322
Service: Dropbox * File Size	Google Drive, N/A	0.020	0.160	0.127	0.899
Service: Dropbox * Shared	Google Drive, Not	0.859	0.611	1.406	0.160

The results of a mixed-effects logistic regression to identify what factors were correlated with **expressing a preference to encrypt the file shown**, as opposed to keeping the file as-is. Files the participant wanted to delete are excluded from this model. Non-italicized values in the baseline column specify the baseline category for terms representing categorical variables. Italicized values in the baseline column indicate the units for numerical terms. Significant p-values are shown in bold.

Table XVI: Appendix - Factors correlated with preferences for file encryption.

Factor	Baseline / <i>Units</i>	Coefficient	Std. Error	z value	p
Service: Dropbox	Google Drive	-4.400	2.808	-1.567	0.117
File Type: Document	Other	-0.348	0.645	-0.539	0.590
File Type: Image	Other	-0.424	0.680	-0.624	0.533
File Type: Spreadsheet	Other	-24.054	420.899	-0.057	0.954
File Type: Video	Other	-1.527	1.340	-1.139	0.255
Access: Editor	Owner	0.255	0.982	0.260	0.795
Access: Viewer	Owner	-23.491	280.600	-0.084	0.933
Days Since Modified	<i>log10(days+1)</i>	-0.034	0.341	-0.099	0.921
File Size	<i>log10(bytes)</i>	-0.256	0.189	-1.355	0.176
Shared	Not shared	-0.086	0.634	-0.135	0.892
Account Age	<i>Years</i>	0.105	0.234	0.451	0.652
Participant Tech. Background	No	1.177	0.562	2.095	<b>0.036</b>
Participant Age	<i>Years</i>	-0.032	0.029	-1.089	0.276
Account for Work Purposes	No	-1.37	0.549	-2.486	<b>0.013</b>
Account for Personal Purposes	No	-0.594	0.624	-0.952	0.341
Service: Dropbox * File Type: Document	Google Drive, Other	1.098	1.228	0.894	0.371
Service: Dropbox * File Type: Image	Google Drive, Other	1.375	1.108	1.241	0.215
Service: Dropbox * File Type: Spreadsheet	Google Drive, Other	28.213	420.896	0.067	0.947
Service: Dropbox * File Type: Video	Google Drive, Other	1.583	1.987	0.797	0.426
Service: Dropbox * Access Type: Editor	Google Drive, Owner	0.156	1.442	0.108	0.914
Service: Dropbox * Access Type: Viewer	Google Drive, Owner	24.414	280.597	0.087	0.931
Service: Dropbox * Days Since Modified	Google Drive, N/A	-0.086	0.571	-0.150	0.881
Service: Dropbox * File Size	Google Drive, N/A	0.517	0.290	1.783	0.075
Service: Dropbox * Shared	Google Drive, Not	0.189	1.056	0.179	0.858

The results of a mixed-effects ordinal regression to identify what factors were correlated with **expressing a preference to stop sharing the file shown**. In particular the dependent variable is an ordinal variable reflecting a preference to keep sharing (1), whether the sharing setting does not matter (2), or to stop sharing (3). Non-italicized values in the baseline column specify the baseline category for terms representing categorical variables. Italicized values in the baseline column indicate the units for numerical terms. Significant p-values are shown in bold.

Table XVII: Appendix - Factors correlated with wanting to stop sharing.

<b>Factor</b>	<b>Baseline / <i>Units</i></b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z value</b>	<b>p</b>
Service: Dropbox	Google Drive	-3.083	1.492	-2.066	<b>0.038</b>
File Type: Document	Other	0.303	1.266	0.240	0.810
File Type: Image	Other	-1.211	1.264	-0.958	0.338
File Type: Spreadsheet	Other	2.470	2.072	1.192	0.232
File Type: Video	Other	-2.945	1.899	-1.551	0.121
Access: Editor	Owner	-0.466	1.009	-0.462	0.643
Access: Viewer	Owner	-1.610	3.936	-0.408	0.683
Days Since Modified	<i>log10(days+1)</i>	1.729	1.028	1.682	0.092
File Size	<i>log10(bytes)</i>	0.861	0.348	2.472	<b>0.013</b>
Account Age	<i>Years</i>	-0.159	0.715	-0.223	0.823
Participant Tech. Background	No	-1.234	1.509	-0.818	0.413
Participant Age	<i>Years</i>	-0.006	0.072	-0.091	0.927
Account for Work Purposes	No	-0.261	1.431	-0.183	0.854
Account for Personal Purposes	No	1.324	1.547	0.856	0.392
Relationship to Sharing Recipient	Have communicated in past year	3.422	0.777	-0.360	<b>&lt;.001</b>

## Cumulative Distributions of Account Level Properties

The cumulative distribution of the age of participants' accounts. As Google Drive started on April 24, 2012, the oldest Google Drive account is around 5 years old.

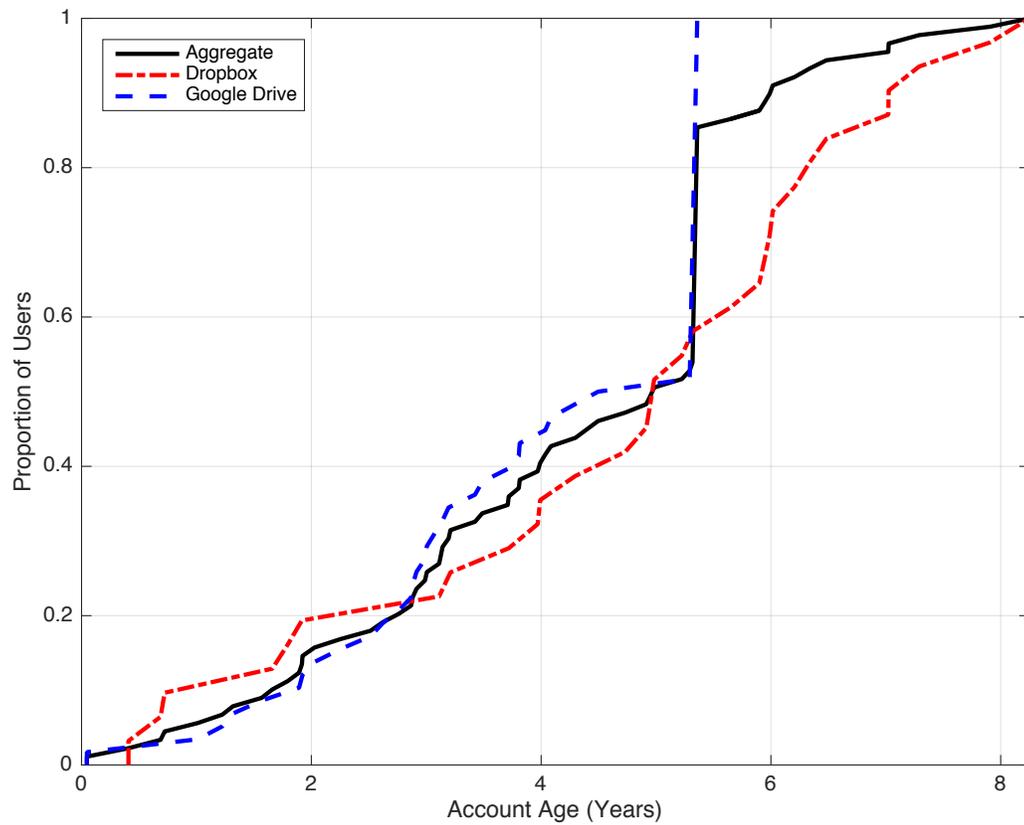


Figure 18: Appendix - The cumulative distribution of account age.

The distribution of how participants' accounts varied in (the natural logarithm of) the number of bytes stored. Dropbox and Google Drive follow similar trends.

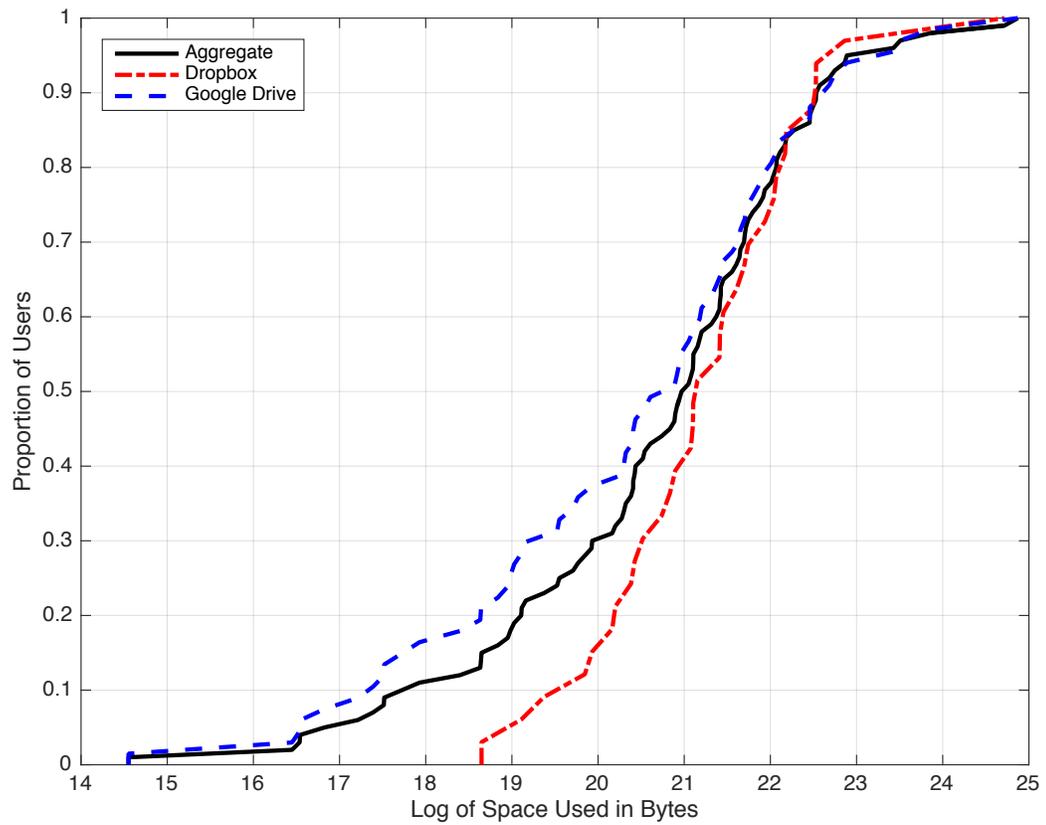


Figure 19: Appendix - The cumulative distribution of account size.

The distribution of (the natural logarithm of) the number of files in participants' accounts. Dropbox and Google Drive follow similar trends.

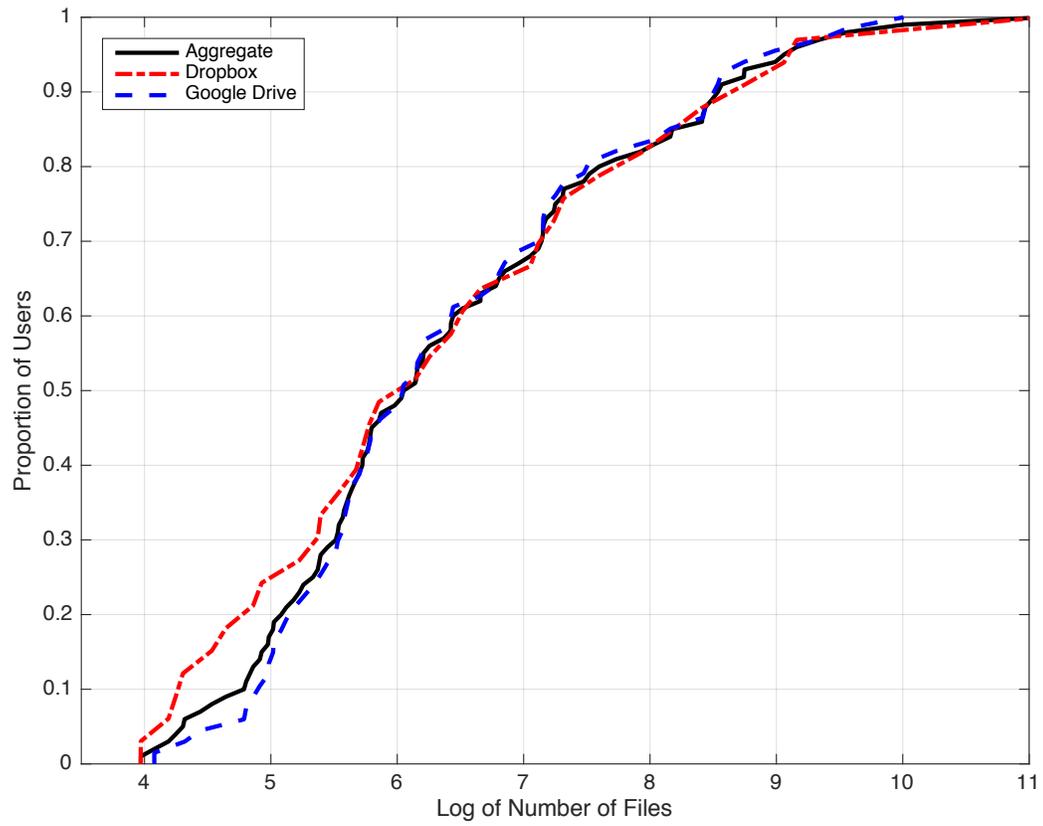


Figure 20: Appendix - The logarithmic distribution of account files.

File sharing trends among Dropbox and Google Drive participants.

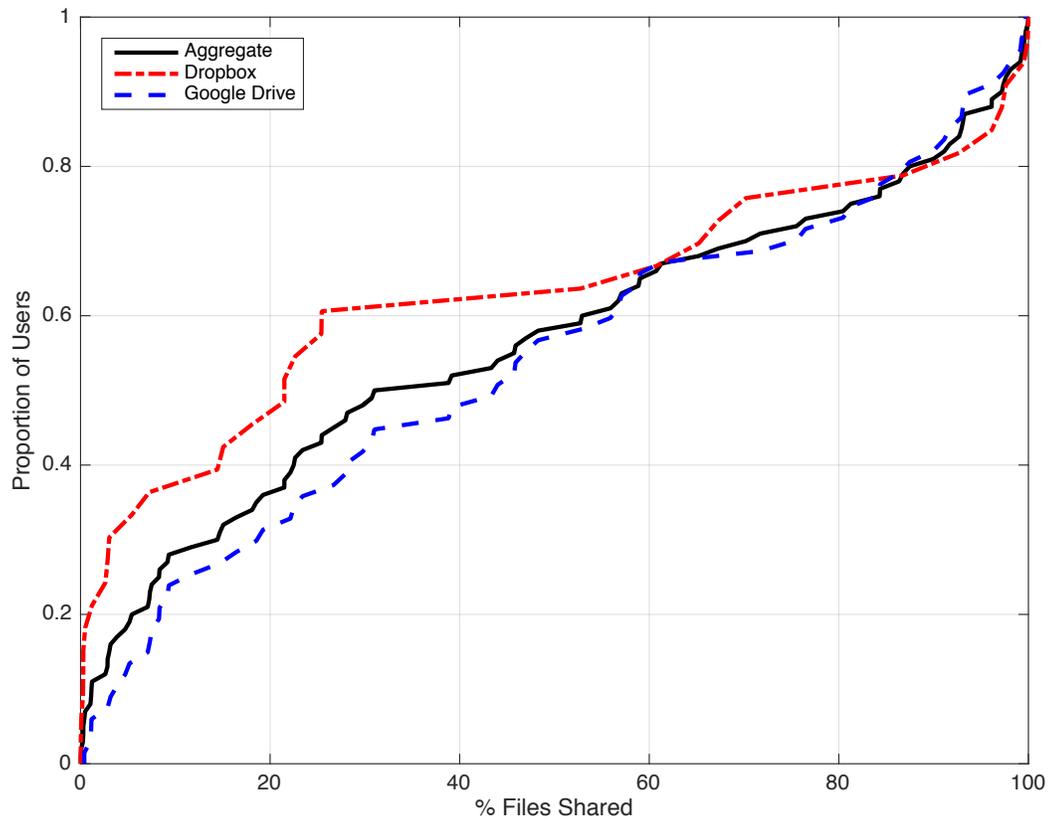


Figure 21: Appendix - The cumulative distribution of shared files.

## Qualitative Interview Script

### Part 1: Generic Questions

#### *Questions on general cloud storage usage*

**Q:** What is useful about storing your files in the cloud?

**Q:** What is difficult about storing your files in the cloud, or about using your files that are stored in the cloud?

#### *Questions on the sensitivity of files in the cloud*

**Q:** In general, what do you feel that makes information or data stored on your cloud sensitive?

**Q:** If you were given the opportunity to list out reasons of why a file is could be sensitive, what would all the reasons be?

**Q:** Can you provide us of any examples of sensitive files you might have online in the cloud?

**Q:** Do you have any files elsewhere which you hesitated to upload online because they were sensitive?

*Next you will be provided with different scenarios regarding the sensitivity of files. You will be required to provide examples of files that you believe to fit the criteria. You can provide file categories, or be more specific by providing file extensions, etc.*

**Q:** Please provide example files in the following scenarios:

- Files you would be concerned about if they were hacked from your cloud.
- Files on the cloud, if which were made public on to social media, that worry or embarrass you
- Files on the cloud which would concern you if your close family members saw them.

#### *Questions on the usefulness of files in the cloud*

**Q:** If you were given the opportunity to list out reasons of why a file is useful, what would all the reasons be?

**Q:** Can you provide us of files which are useful to you have online or elsewhere?

*Next you will be provided with different scenarios regarding the usefulness of files. You will be required to provide examples of files that you believe to fit the criteria. You can provide file categories, or be more specific by providing file extensions, etc.*

**Q:** Please provide example files in the following scenarios:

- Files you would recover if they were accidentally deleted from your cloud account
- Files on the cloud that you access and update on a regular basis.
- Files on the cloud that you have shared with your friends and family

## **Part 2: File-Specific Questions**

### *Sensitivity-specific questions*

**Q:** Do you consider this file to be sensitive?

*Only ask this question if there was an affirmative response to “do you consider this file to be sensitive?”*

**Q:** What particularly do you consider sensitive about this file?

**Q:** Would you be concerned if this file was breached in an event of a hack or unauthorized access to your account?

**Q:** How would you feel about protecting this file in a way such that every time you access it, you will be required to input a password or a secret key

### *Usefulness-specific questions*

**Q:** Is the file shown currently useful to you?

**Q:** Will this file be useful to you in the future?

*Only ask this question if there was an affirmative response to “is the file shown currently useful to you?” OR an affirmative response to “will this file be useful to you in the future?”*

**Q:** What particularly do you consider sensitive about this file?

**Q:** Would you like to keep this file forever on the cloud?

## Data Collection Visualization

### What we collect...

#### Metadata Collected

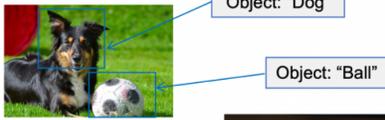
- Encrypted file names
- File size in bytes
- File type and extension
- Encrypted names of folders containing file
- File creation, access, share and modification times
- Encrypted names of file owners and shared users

#### File Data Collected

- Linguistic context of words in documents
- Topic associations of documents
- Numerical data encoding distribution of words in documents
- Image dimensions
- Image average coloration and distribution of coloration

#### Additional Image Data Collected

The examples demonstrate the image related features we collect about using the Google Vision API




#### We **DO NOT** Collect:

- Usernames
- Passwords
- Email addresses
- Original file contents
- Any other personally identifiable information

Figure 22: Appendix - Description of features collected from cloud files.

## Quantitative Study Survey Instrument

Questions prefixed with the  $\Rightarrow$  symbol indicate conditional-branched questions.

### 1. Generic Questions

**G-1 For approximately how long have you had the *Cloud Storage* account you are using for this study?**

- Less than 1 year
- At least 1 year, but less than 2 years
- At least 2 years, but less than 3 years
- At least 3 years, but less than 4 years
- At least 4 years, but less than 5 years
- More than 5 years

**G-2 Cloud storage providers offer both free accounts and paid accounts, where the latter offers more storage space. Do you use a free *Cloud Storage* account or a paid *Cloud Storage* account?**

- Free account
- I pay for it
- My school/work pays for it
- Don't know

**G-3 Which best describes how often you use this *Cloud Storage* account for work or school purposes?**

- At least once a week
- At least once a month, but less than once a week
- At least once a year, but less than once a month
- Less than once a year, but sometimes
- I do not use it for work or school purposes
- Don't Know

**G-4 Which best describes how often you use this *Cloud Storage* account for personal purposes (i.e., for purposes other than for work or school)?**

- At least once a week
- At least once a month, but less than once a week
- At least once a year, but less than once a month
- Less than once a year, but sometimes
- I do not use it for work or school purposes
- Don't Know

**G-5 For which tasks below do you use this *Cloud Storage* account? (Check all that apply)**

- Collaborating with co-workers, classmates, or professional contacts by jointly creating and editing files
- Collaborating with friends and family by jointly creating and editing files
- Sharing files that I have created with co-workers, classmates, or other professional contacts
- Sharing files that I have created with family and friends
- Backing up files related to my job, school, or career
- Backing up files that are related not related to my job, school, or career

Other

**G-6** There are multiple ways you can access files in your *Cloud Storage* account. One is to access files in your *Cloud Storage* account is by using a web browser like Chrome, Firefox, or Safari to log into the *Cloud Storage* website.

- Daily or more frequently
- Every few days
- Weekly
- Monthly
- Less than once a month, but sometimes
- Never
- Don't Know

**G-7** Another way to access files in your *Cloud Storage* account is by using an app on your smartphone (iPhone or Android). How often do you use a smartphone app to access files or folders stored in this *Cloud Storage* account?

- Daily or more frequently
- Every few days
- Weekly
- Monthly
- Less than once a month, but sometimes
- Never
- Don't Know

**G-8** Another way is by installing *Cloud Storage* software on your computer so that certain folders are automatically synced with your *Cloud Storage* account. How often do you access (view or edit) files or folders on your computer that are automatically synced with your *Cloud Storage* account?

- Daily or more frequently
- Every few days
- Weekly
- Monthly
- Less than once a month, but sometimes
- Never
- Don't Know

*For the next question, we define media files mainly as images, videos and audio files. Documents are text-based files such as word documents, pdfs, ebooks, spreadsheets and slides. All other files belong to the other category.*

**G-9 Among these types of files, which do you most commonly store in your *Cloud Storage* account?**

- Mostly media files and only a few documents or other files
- Mostly documents and only a few media or other files
- Mostly other files and only a few media files or documents
- Files of all types, with no clear majority
- Other
- Don't Know

**G-10 Which best describes how often you revisit your *Cloud Storage* to organize it by deleting unnecessary files, moving files to different folders, or performing similar clean-up tasks?**

- At least once a week
- At least once a month
- At least once a year
- Less than once a year
- I have never organized my *Cloud Storage*
- Don't know

**G-11 Please state your agreement with the following statement: My *Cloud Storage* account is organized.**

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

**G-12 Please spend at least a minute answering the following question: What are the different reasons you might want to keep a file in your *Cloud Storage*?**

**G-13 Do you believe that you have files in your *Cloud Storage* that you may need to reference or use in the future?**

- Yes
- No
- Don't know

⇒ **G-131 Please describe what kinds of files in are your *Cloud Storage* that you may need to reference or use in the future?**

⇒ **G-132 Why don't you have any such files in your *Cloud Storage*?**

**G-14** Do you anticipate you have files in your *Cloud Storage* that have sentimental value?

- Yes
- No
- Don't know

⇒ **G-141** What are the kinds of files on your *Cloud Storage* that have sentimental value?

⇒ **G-142** Why don't you have any such files in your *Cloud Storage*?

**G-15** Do you anticipate you have files in your *Cloud Storage* that are for backup purposes?

- Yes
- No
- Don't know

⇒ **G-151** What are the kinds of files on your *Cloud Storage* that are for backup purposes?

⇒ **G-152** Why don't you have any such files in your *Cloud Storage*?

**G-16** Please spend at least a minute answering the following question: What are the different reasons you might consider a file in your *Cloud Storage* potentially harmful, risky, or dangerous if it were accessed by an unauthorized party?

**G-17** Do you anticipate you have files in your *Cloud Storage* that may contain personally identifiable or financial information about you?

- Yes
- No
- Don't know

⇒ **G-171** What are the kinds of files in your *Cloud Storage* that may contain personally identifiable or financial information about you?

⇒ **G-172** Why don't you have any such files in your *Cloud Storage*?

**G-18** Do you anticipate you have files in your *Cloud Storage* that may contain personally identifiable or financial information about people other than you?

- Yes
- No
- Don't know

⇒ **G-181** What are the kinds of files in your *Cloud Storage* that may contain personally identifiable or financial information about people other than you?

⇒ **G-182** Why don't you have any such files in your *Cloud Storage*?

**G-19** Do you believe that you have files in your *Cloud Storage* that may contain intimate or embarrassing content?

- Yes
- No
- Don't know

⇒ **G-191** What are the kinds of files in your *Cloud Storage* that may contain intimate or embarrassing content?

⇒ **G-192** Why don't you have any such files in your *Cloud Storage*?

**G-20** Do you believe that you have files in your *Cloud Storage* that may contain content that you have created (e.g., art, writing)?

- Yes
- No
- Don't know

⇒ **G-201** What are the kinds of files in your *Cloud Storage* that may contain that may contain content that you have created (e.g., art, writing)?

⇒ **G-202** Why don't you have any such files on your *Cloud Storage*?

**G-21** Do you believe that you have files in your *Cloud Storage* that may contain proprietary information (e.g., from your workplace)?

- Yes
- No
- Don't know

⇒ **G-211** What are the kinds of files in your *Cloud Storage* that may contain proprietary information (e.g., from your workplace)?

⇒ **G-212** Why don't you have any such files on your *Cloud Storage*?

**G-22** How concerned would you be if the files stored in this *Cloud Storage* account were leaked in a data breach?

- Extremely concerned
- Moderately concerned
- Somewhat concerned
- Slightly concerned
- Not at all concerned

**G-23** Have you enabled two-factor authentication on your account (in which you use your phone/email for verification in addition to your password)?

- Yes
- No
- Don't know

**G-24** Have you taken any additional steps to protect files in your account from a potential data breach?

- Yes
- No
- Don't know

⇒ **G-241** What kind of steps have you taken to protect your files from a data breach? \*

⇒ **G-242** Why have you not taken any steps to protect your files from a data breach?

## 2. File-Specific Questions

**U-1 Please rate your agreement with this statement: “I consider this file worth keeping.”**

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

⇒ **U-11 Why is this file worth keeping? (Select all that apply)**

- I may need to reference this file in the future
- This file has sentimental value associated with it
- The file is useful for backup purposes
- Other

⇒ **U-12 Why is this file not worth keeping? (Select all that apply)**

- This file is no longer useful
- I have a copy of this file elsewhere
- I do not recognize this file
- I do not remember originally storing this file to my *Cloud Storage*
- It would be risky to keep this file
- Other

⇒ **U-13 How long do you expect this file to be worth keeping?**

- Forever
- For at least 5 years, but not forever
- For the next 1 - 5 years
- For at most the next year

**U-2 To the best of your knowledge, when was the last time you accessed (viewed or modified) this file?**

- Within the last month
- Between one month and one year ago
- Between one year and five years ago
- At least 5 years ago
- Never

**U-3 When do you next expect to access (view or modify) this file in the future?**

- Within the last month
- Between one month and one year ago
- Between one year and five years ago
- Over 5 years from now, but eventually
- Never

**S-1 Please rate your agreement with this statement: “It would be risky, harmful, or otherwise dangerous if this file were accessed without my consent.”**

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

⇒ **S-11 Why is this file risky, harmful, or dangerous? (Select all that apply)**

- It contains personally identifiable information or financial information
- It contains intimate or embarrassing content
- It contains intellectual property or has proprietary information
- Other

⇒ **S-12 Why is this file not risky, harmful, or dangerous? (Select all that apply)**

- It contains information that is already public
- It contains content that does need to be protected
- Other

*How would you like to manage this file from among the three options listed below?*

1. *Keep As-Is: The file will remain in your cloud storage account in its current state.*
2. *Delete: The file will be removed from your cloud storage account*
3. *Protect: The file will remain in your cloud storage account. However, you will need to take extra security steps to access the contents of the file.*

**M-1 Which of these statements best characterizes what you would like to happen to this file?**

- I would like to keep this file stored as-is in my *Cloud Storage*
- I would like to delete this file from my *Cloud Storage*
- I would like to protect this file in my *Cloud Storage*

⇒ **M-11 Why would you want to keep this file as-is in your *Cloud Storage*?**

⇒ **M-12 Why would you want to delete this file from your *Cloud Storage*?**

⇒ **M-13 Why would you want to keep a protected version of this file in your *Cloud Storage*?**

### 3. Demographic Questions

**D-1 With what gender do you identify?**

- Male
- Female
- Non-binary
- Other
- Prefer not to answer

**D-2 Are you majoring in, or do you have a degree or job in, any of the following fields?: computer science; computer engineering; information technology; or a related field?**

- Yes
- No
- Prefer not to answer

**D-3 How old are you? (optional)**

**D-4 What is your occupation?**

UNIVERSITY OF ILLINOIS  
AT CHICAGO

Office for the Protection of Research Subjects (OPRS)  
Office of the Vice Chancellor for Research (MC 672)  
203 Administrative Office Building  
1737 West Polk Street  
Chicago, Illinois 60612-7227

**Exemption Granted**

February 23, 2017

Chris Kanich, Ph.D.  
Computer Science  
851 S Morgan Street, SEO1120  
M/C 152  
Chicago, IL 60612  
Phone: (312) 355-0950 / Fax: (312) 413-0024

**RE: Research Protocol # 2017-0186**  
**“Characterizing Cloud Storage: Use Cases, Latent Risks, and Opportunities for Improvement”**

**Sponsors: None**

**Please note the following:**

1. Please be reminded of the need to address University of Chicago approval requirements.
2. Please have Mohammad Taha Khan complete Investigator Continuing Education prior to conducting this research. His UIC Investigator Training Period expired on January 16, 2017.
3. Please ensure the following University of Chicago key research personnel meet institutional investigator training requirements: Maia Hyun and Blase Ur.

Dear Dr. Kanich:

Your Claim of Exemption was reviewed on February 23, 2017 and it was determined that your research protocol meets the criteria for exemption as defined in the U. S. Department of Health and Human Services Regulations for the Protection of Human Subjects [(45 CFR 46.101(b)]. You may now begin your research.

**UIC Exemption Period:** February 23, 2017 – February 23, 2020  
**Lead Performance Site:** UIC  
**Other Performance Site(s):** University of Chicago (see text box above)  
**Subject Population:** Adult (18+ years) subjects only  
**Number of Subjects:** 150

**The specific exemption category under 45 CFR 46.101(b) is:**

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

You are reminded that investigators whose research involving human subjects is determined to be exempt from the federal regulations for the protection of human subjects still have responsibilities for the ethical conduct of the research under state law and UIC policy. Please be aware of the following UIC policies and responsibilities for investigators:

1. Amendments You are responsible for reporting any amendments to your research protocol that may affect the determination of the exemption and may result in your research no longer being eligible for the exemption that has been granted.
2. Record Keeping You are responsible for maintaining a copy all research related records in a secure location in the event future verification is necessary, at a minimum these documents include: the research protocol, the claim of exemption application, all questionnaires, survey instruments, interview questions and/or data collection instruments associated with this research protocol, recruiting or advertising materials, any consent forms or information sheets given to subjects, or any other pertinent documents.
3. Final Report When you have completed work on your research protocol, you should submit a final report to the Office for Protection of Research Subjects (OPRS).
4. Information for Human Subjects UIC Policy requires investigators to provide information about the research to subjects and to obtain their permission prior to their participating in the research. The information about the research should be presented to subjects as detailed in the research protocol and application utilizing the approved recruitment and consent process and document(s).

Please be sure to use your research protocol number (listed above) on any documents or correspondence with the IRB concerning your research protocol.

We wish you the best as you conduct your research. If you have any questions or need further help, please contact me at (312) 355-2908 or the OPRS office at (312) 996-1711.

Sincerely,  
Charles W. Hoehne, B.S., C.I.P.  
Assistant Director, IRB #7  
Office for the Protection of Research Subjects

cc: Robert Sloan, Computer Science, M/C 152



**Exemption Determination  
Amendment to Research Protocol – Exempt Review  
UIC Amendment #1**

October 31, 2017

Chris Kanich, Ph.D.  
Computer Science  
851 S Morgan Street, SEO1120  
M/C 152  
Chicago, IL 60612  
Phone: (312) 355-0950 / Fax: (312) 413-0024

**RE: Protocol # 2017-0186  
“Characterizing Cloud Storage: Use Cases, Latent Risks, and Opportunities for Improvement”**

Dear Dr. Kanich:

The OPRS staff/members of Institutional Review Board (IRB) #7 have reviewed this amendment to your research, and have determined that your research protocol continues to meet the criteria for exemption as defined in the U. S. Department of Health and Human Services Regulations for the Protection of Human Subjects [(45 CFR 46.101(b)).

**The specific exemption category under 45 CFR 46.101(b) is:**

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

You may now implement the amendment in your research.

Please note the following information about your approved amendment:

**Exemption Period: October 31, 2017 – October 31, 2020**

**Amendment Approval Date: October 31, 2017**

**Amendment:**

Summary: UIC Amendment #1: The principal investigator is adding a new faculty member, Elena Zheleva, and her research assistant, Christopher Tran, to the protocol so that they may assist in performing the data analysis for this project. The data collection phase of this research has been completed.

You are reminded that investigators whose research involving human subjects is determined to be exempt from the federal regulations for the protection of human subjects still have responsibilities for the ethical conduct of the research under state law and UIC policy. Please be aware of the



following UIC policies and responsibilities for investigators:

1. Amendments You are responsible for reporting any amendments to your research protocol that may affect the determination of the exemption and may result in your research no longer being eligible for the exemption that has been granted.
2. Record Keeping You are responsible for maintaining a copy all research related records in a secure location in the event future verification is necessary, at a minimum these documents include: the research protocol, the claim of exemption application, all questionnaires, survey instruments, interview questions and/or data collection instruments associated with this research protocol, recruiting or advertising materials, any consent forms or information sheets given to subjects, or any other pertinent documents.
3. Final Report When you have completed work on your research protocol, you should submit a final report to the Office for Protection of Research Subjects (OPRS).

Please be sure to use your research protocol number (2017-0186) on any documents or correspondence with the IRB concerning your research protocol.

We wish you the best as you conduct your research. If you have any questions or need further help, please contact me at (312) 355-2908 or the OPRS office at (312) 996-1711.

Sincerely,  
Charles W. Hoehne, B.S., C.I.P.  
Assistant Director, IRB #7  
Office for the Protection of Research Subjects

cc: Robert Sloan, Computer Science, M/C 152



**Exemption Determination  
Amendment to Research Protocol – Exempt Review  
UIC Amendment #2**

February 27, 2018

Chris Kanich, Ph.D.  
Computer Science  
851 S Morgan Street, SEO1120, M/C 152  
Chicago, IL 60612  
Phone: (312) 355-0950 / Fax: (312) 413-0024

**RE: Protocol # 2017-0186  
“Characterizing Cloud Storage: Use Cases, Latent Risks, and Opportunities for  
Improvement”**

**Please note: Although the reviewers have approved this amendment, they have determined Will Brackenbury, University of Chicago student who will be helping with the analysis of data collected in such a manner that individuals cannot be directly or indirectly identified, does NOT meet the institutional definition of key research personnel:**

**Key Research Personnel** include all persons who will have a significant role in the design or conduct of the research, and includes at a minimum all Principal Investigators and Co-Investigators, and any individuals who are individually named on a grant or contract application, who are listed on an FDA form 1572 (for the conduct of the research at UIC), who are named as contact persons in the informed consent documents or recruitment materials for research, and persons who are, or who provide supervision of the persons who are, obtaining informed consent to participate in research.

Additionally, any individuals (including student researchers and coordinators) who are involved with the research by handling protected health information or are using the research information/data set as part of their own research should be included as research personnel on a protocol application.

If students or other individuals have minor roles in the research that are not listed above, they are not required to be listed on the research protocol. However, the Principal Investigator is responsible to ensure that these individuals receive both adequate training, including human subjects protection training, and oversight in accordance to the roles these individuals perform in the research.

Dear Dr. Kanich:

The OPRS staff/members of Institutional Review Board (IRB) #7 have reviewed this amendment to your research, and have determined that your amended research continues to meet the criteria for exemption as defined in the U. S. Department of Health and Human Services Regulations for the Protection of Human Subjects [(45 CFR 46.101(b)).



**The specific exemption category under 45 CFR 46.101(b) is: 2**

You may now implement the amendment in your research.

Please note the following information about your approved amendment:

**Exemption Period:** February 27, 2018 – February 27, 2021

**Amendment Approval Date:** February 27, 2018

**Amendment:**

Summary: UIC Amendment #2: Addition of Will Brackenbury, University of Chicago student, as a research assistant to conduct data analysis.

You are reminded that investigators whose research involving human subjects is determined to be exempt from the federal regulations for the protection of human subjects still have responsibilities for the ethical conduct of the research under state law and UIC policy. Please be aware of the following UIC policies and responsibilities for investigators:

1. **Amendments** You are responsible for reporting any amendments to your research protocol that may affect the determination of the exemption and may result in your research no longer being eligible for the exemption that has been granted.
2. **Record Keeping** You are responsible for maintaining a copy all research related records in a secure location in the event future verification is necessary, at a minimum these documents include: the research protocol, the claim of exemption application, all questionnaires, survey instruments, interview questions and/or data collection instruments associated with this research protocol, recruiting or advertising materials, any consent forms or information sheets given to subjects, or any other pertinent documents.
3. **Final Report** When you have completed work on your research protocol, you should submit a final report to the Office for Protection of Research Subjects (OPRS).
4. **Information for Human Subjects** UIC Policy requires investigators to provide information about the research to subjects and to obtain their permission prior to their participating in the research. The information about the research should be presented to subjects as detailed in the research protocol, application and supporting documents.

Please be sure to use your research protocol number (2017-0186) on any documents or correspondence with the IRB concerning your research protocol.

We wish you the best as you conduct your research. If you have any questions or need further



help, please contact me at (312) 355-2908 or the OPRS office at (312) 996-1711.

Sincerely,  
Charles W. Hoehne, B.S., C.I.P.  
Assistant Director, IRB #7  
Office for the Protection of Research Subjects

cc: Robert Sloan, Computer Science, M/C 152



**Exemption Determination  
Amendment to Research Protocol – Exempt Review  
UIC Amendment #3**

June 8, 2018

Chris Kanich, Ph.D.  
Computer Science  
Phone: (312) 355-0950 / Fax: (312) 413-0024

**RE: Protocol # 2017-0186  
“Characterizing Cloud Storage: Use Cases, Latent Risks, and Opportunities for  
Improvement”**

**Approval of UIC Amendment #3 does not include approval of Mainack Mondal as co-investigator/key research personnel.** To add Mainack Mondal as KRP, please submit a separate amendment and include:

1. A copy of Mainack Mondal's initial investigator training certificate; and
2. A copy of the University of Chicago IRB approval letter or exemption determination.

Dear Dr. Kanich:

The OPRS staff/members of Institutional Review Board (IRB) #7 have reviewed this amendment to your research, and have determined that your research protocol continues to meet the criteria for exemption as defined in the U. S. Department of Health and Human Services Regulations for the Protection of Human Subjects [(45 CFR 46.101(b)].

**The specific exemption category under 45 CFR 46.101(b) is: 2**

You may now implement the amendment in your research.

Please note the following information about your approved amendment:

**UIC Exemption Period: June 8, 2018 – June 7, 2021**

**Amendment Approval Date: June 8, 2018**

**Amendment:**

Summary: UIC Amendment #3: Principal Investigator's summary:

1. We are adding a funding source (a NSF grant).
2. We will select the files to present to users for closer investigation based on a machine learning algorithm. This algorithm will take as input non-PII, non-sensitive file metadata as input, including both content metadata (which is already collected in the approved version of the protocol), as well as high level labels (e.g, person, boat, document), and choose files which are most likely to need reevaluation by the user.
3. We modify the survey to ask more precise questions about the perceived utility and sensitivity of old files for selecting content to identify actionable factors behind retrospective file management in cloud storage.
4. We will launch the research study to another set of participants (distinct from the



participants that have taken part in the earlier study), which will necessitate an increase in the total number of research participants to 500.

You are reminded that investigators whose research involving human subjects is determined to be exempt from the federal regulations for the protection of human subjects still have responsibilities for the ethical conduct of the research under state law and UIC policy. Please be aware of the following UIC policies and responsibilities for investigators:

1. Amendments You are responsible for reporting any amendments to your research protocol that may affect the determination of the exemption and may result in your research no longer being eligible for the exemption that has been granted.
2. Record Keeping You are responsible for maintaining a copy all research related records in a secure location in the event future verification is necessary, at a minimum these documents include: the research protocol, the claim of exemption application, all questionnaires, survey instruments, interview questions and/or data collection instruments associated with this research protocol, recruiting or advertising materials, any consent forms or information sheets given to subjects, or any other pertinent documents.
3. Final Report When you have completed work on your research protocol, you should submit a final report to the Office for Protection of Research Subjects (OPRS).
4. Information for Human Subjects UIC Policy requires investigators to provide information about the research to subjects and to obtain their permission prior to their participating in the research. The information about the research should be presented to subjects as detailed in the research protocol, application and supporting documents.

Please be sure to use your research protocol number (2017-0186) on any documents or correspondence with the IRB concerning your research protocol.

We wish you the best as you conduct your research. If you have any questions or need further help, please contact me at (312) 355-2908 or the OPRS office at (312) 996-1711.

Sincerely,  
Charles W. Hoehne  
Assistant Director, IRB #7  
Office for the Protection of Research Subjects

cc: Robert Sloan, Computer Science, M/C 152

## ACM Permission and Release Form

Title of non-ACM work: Forgotten But Not Gone: Identifying the Need for Longitudinal Data Management in Cloud Storage Submission ID: **pn4337**

Author(s): Mohammad Taha Khan (Univ. of Illinois at Chicago); Maria Hyun (Univ. of Chicago); Chris Kanich (Univ. of Illinois at Chicago); Blase Ur (Univ. of Chicago)

Type of material: **Full Paper; supplemental material(s)**

TITLE OF ACM PUBLICATION: CHI 2018: CHI Conference on Human Factors in Computing Systems Proceedings

### Grant Permission

As the owner or authorized agent of the copyright owner(s) I hereby grant non-exclusive permission for ACM to include the above-named material (the *Material*) in any and all forms, in the above-named publication.

I further grant permission for ACM to distribute or sell this submission as part of the above-named publication in electronic form, and as part of the ACM Digital Library, compilation media (CD, DVD, USB) or broadcast, cablecast, laserdisc, multimedia or any other media format now or hereafter known. (*Not all forms of media will be utilized.*)

You have opted to pay an article processing fee in exchange for permanent OA (Open Access) for your article in the ACM Digital Library. Your article will become open upon receipt of payment. Remember that your choice to retain copyright and grant ACM non-exclusive permission to publish is conditional on your pledge to make payment for permanent open access.

Yes, I grant ACM permission as stated above and I further I agree that if I fail to make full payment of the article processing fee within 6 months of the article publication date, ACM may automatically convert my grant of non-exclusive permission to publish to the ACM Exclusive Publishing License.

### Multiple Author Submission Options

- I am submitting this permission and release form on behalf of all co-authors  
 I cannot submit this permission and release form on behalf of all co-authors

The following notice of publication and ownership will be displayed with the Material in all publication formats:

The new [ACM Consolidated TeX template Version 1.3 and above](#) automatically creates and positions these text blocks for you based on the code snippet which is system-generated based on your rights management choice and this particular conference.

*Please copy and paste the following code snippet into your TeX file between `\begin{document}` and `\maketitle`, either after or before CCS codes.*

```
\copyrightyear{2018}
\acmYear{2018}
\setcopyright{rightsretained}
\acmConference[CHI 2018]{CHI Conference on Human Factors in
Computing Systems }{April 21--26, 2018}{Montreal, QC, Canada}
\acmBooktitle{CHI 2018: CHI Conference on Human Factors in Computing
Systems , April 21--26, 2018, Montreal, QC, Canada}
\acmDOI{10.1145/3173574.3174117}
\acmISBN{978-1-4503-5620-6/18/04}
```

ACM TeX template .cls version 2.8, automatically creates and positions these text blocks for you based on the code snippet which is system-generated based on your rights management choice and this particular conference.

*Please copy and paste the following code snippet into your TeX file between `\begin{document}` and `\maketitle`, either after or before CCS codes.*

```
\CopyrightYear{2018}
\setcopyright{rightsretained}
\conferenceinfo{CHI 2018}{April 21--26, 2018, Montreal, QC,
Canada}\isbn{978-1-4503-5620-6/18/04}
\doi{https://doi.org/10.1145/3173574.3174117}
```

*If you are using the ACM Microsoft Word template, or still using an older version of the ACM TeX template, or the current versions of the ACM SIGCHI, SIGGRAPH, or SIGPLAN TeX templates, you must copy and paste the following text block into your document as per the instructions provided with the templates you are using:*

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada  
© 2018 Copyright is held by the owner/author(s).  
ACM ISBN 978-1-4503-5620-6/18/04.  
<https://doi.org/10.1145/3173574.3174117>

#### **Audio/Video Release**

\* Your Audio/Video Release is conditional upon you agreeing to the terms set out below.

I further grant permission for ACM to record and/or transcribe and reproduce my

presentation and likeness in the conference publication and as part of the ACM Digital Library and to distribute the same for sale in complete or partial form as part of an ACM product on CD-ROM, DVD, webcast, USB device, streaming video or any other media format now or hereafter known.

I understand that my presentation will not be sold separately as a stand-alone product without my direct consent. Accordingly, I further grant permission for ACM to include my name, likeness, presentation and comments and any biographical material submitted by me in connection with the conference and/or publication, whether used in excerpts or in full, for distribution described above and for any associated advertising or exhibition.

Do you agree to the recording, transcription and distribution?  Yes  No

#### **Auxiliary Materials, not integral to the Work**

Do you have any Auxiliary Materials?  Yes  No

\* Your Auxiliary Materials Release is conditional upon you agreeing to the terms set out below.

[Defined as additional files, video or software and executables that are not submitted for review and publication as an integral part of the Work but are supplied by the author as useful resources.] I hereby grant ACM permission to serve files containing my Auxiliary Material from the ACM Digital Library. I hereby represent and warrant that my Auxiliary (software) does not knowingly and surreptitiously incorporate malicious code, virus, trojan horse or other software routines or hardware components designed to permit unauthorized access or to disable, erase or otherwise harm any computer systems or software.

I agree to the above Auxiliary Materials permission statement.

This software is knowingly designed to illustrate technique(s) intended to defeat a system's security. The code has been explicitly documented to state this fact.

#### **Third Party Materials** \* <http://www.acm.org/publications/third-party-material>

In the event that any materials used in my submission or Auxiliary Materials contain the work of third-party individuals or organizations (including copyrighted music or movie excerpts or anything not owned by me), I understand that it is my responsibility to secure any necessary permissions and/or licenses for print and/or digital publication, and cite or attach them below. Third-party copyright must be clearly stated in the caption(s) or images or in the text narrative near the object(s) in the Work and in any presentation of it and in Auxiliary Materials as applicable.

ACM offers Fair Use Guidelines at  
<http://www.acm.org/publications/guidance-for-authors-on-fair-use>

\* Small-performing rights licenses must be secured for the public performance of any copyrighted musical composition. Synchronization licenses must be secured to include any copyrighted musical composition in film or video presentations.

- I have not used third-party material.
- I have used third-party materials and have necessary permissions.

---

---

### **Representations, Warranties and Covenants**

The undersigned hereby represents, warrants and covenants as follows:

- (a) Owner is the sole owner or authorized agent of Owner(s) of the Work;
- (b) The undersigned is authorized to enter into this Agreement and grant the rights included in this license to ACM;
- (c) The Work is original and does not infringe the rights of any third party; all permissions for use of third-party materials consistent in scope and duration with the rights granted to ACM have been obtained, copies of such permissions have been provided to ACM, and the Work as submitted to ACM clearly and accurately indicates the credit to the proprietors of any such third-party materials (including any applicable copyright notice), or will be revised to indicate such credit.
- (d) The Work has not been published except for informal postings on non-peer reviewed servers, and Owner covenants to use best efforts to place ACM DOI pointers on any such prior postings;
- (e) The Auxiliary Materials, if any, contain no malicious code, virus, trojan horse or other software routines or hardware components designed to permit unauthorized access or to disable, erase or otherwise harm any computer systems or software; and
- (f) The Artistic Images, if any, are clearly and accurately noted as such (including any applicable copyright notice) in the Submitted Version.

Additionally, please reference the following representations that must be agreed to prior to submission and acceptance of your paper.

[http://www.acm.org/publications/policies/author\\_representations](http://www.acm.org/publications/policies/author_representations)

- I agree to the Representations, Warranties and Covenants.

DATE: **12/14/2017** sent to blase@uchicago.edu at **23:12:54**

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).

*CHI 2018, April 21–26, 2018, Montréal, QC, Canada.*

ACM ISBN 978-1-4503-5620-6/18/04.

<http://dx.doi.org/10.1145/3173574.3174117>