Forgotten But Not Gone: Identifying the Need for Longitudinal Data Management in Cloud Storage

Mohammad Taha Khan*, **Maria Hyun**†, **Chris Kanich***, **Blase Ur**†
University of Chicago (†) and University of Illinois at Chicago (∗)
taha@cs.uic.edu, mhyun@uchicago.edu, ckanich@uic.edu, blase@uchicago.edu

ABSTRACT
Users have accumulated years of personal data in cloud storage, creating potential privacy and security risks. This agglomeration includes files retained or shared with others simply out of momentum, rather than intention. We presented 100 online-survey participants with a stratified sample of 10 files currently stored in their own Dropbox or Google Drive accounts. We asked about the origin of each file, whether the participant remembered that file was stored there, and, when applicable, about that file's sharing status. We also recorded participants' preferences moving forward for keeping, deleting, or encrypting those files, as well as adjusting sharing settings. Participants had forgotten that half of the files they saw were in the cloud. Overall, 83% of participants wanted to delete at least one file they saw, while 13% wanted to unshare at least one file. Our combined results suggest directions for retrospective cloud data management.

ACM Classification Keywords
H.5.2 User Interfaces: Evaluation/methodology

Author Keywords
Cloud Storage; Dropbox; Google Drive; Longitudinal; Personal Information Management; Deletion; File Sharing

INTRODUCTION
As cloud platforms for storage and backup have matured, many users have implicitly become long-term users of these platforms. These users have years of their personal data stored in the cloud, yet they have likely forgotten about the existence of most of this data. This state of affairs has two troubling consequences. First, the agglomeration of a user's personal data in one location presents attackers with a very attractive target. Compared to the distributed nature of laptops and mobile phones, cloud storage providers are a single point of attack. If an attacker successfully impersonates the user (e.g., by guessing his or her password) or finds a flaw in the cloud implementation (e.g., Apple iCloud had a flaw that allowed unlimited password guessing [14]), the attacker can access potentially all of the user's data. Second, maintaining this large amount of data such that all of it is accessible to the user on a moment's notice is a tremendous waste of resources. As a result, one should aim to remove files that are both risky and useless from the cloud.

In this paper, we conduct a user study to characterize the data participants have stored in their cloud accounts and investigate three types of remediations for retrospective data management: deleting old data, automatically encrypting old data, and moving old data to low-energy archives. To that end, we conducted a 100-participant online-survey using Amazon's Mechanical Turk. To ground this survey concretely in participants' own data, the survey centers around questions we asked about ten files selected from the participant's own Dropbox or Google Drive in a stratified sample. We use the APIs for Dropbox and Google Drive to show participants these files, as well as to characterize their accounts more broadly.

Our survey consisted of three parts. The first part asked generic questions related to account information, such as account age and the main reason for using cloud storage. Second, we asked questions related to the ten different files selected from the user's account, investigating whether participants knew what the file was, whether they remembered that it was stored in the cloud, and gauging whether they wanted to keep the file as-is, or to either delete it or encrypt it. If the file was shared with other users, either by name or via a shared link, we also asked about the origin of this sharing, as well as whether sharing the file was still desired. Finally, we asked about user demographics and general preferences related to the possibility of automated retrospective file management.

Our participants used either Google Drive or Dropbox for storing and sharing a non-trivial number of files, and they had varied goals in using these services: 71% used cloud storage for collaboration, 83% for sharing, and 92% for file archival. Overall, we found that participants' cloud storage accounts contained a mass of data that was indeed forgotten, but not gone. For 51% of the files they saw in the study, participants remembered that the file was stored in the cloud. For 14% of the files they saw, participants did not recognize the file. For the remaining 36%, participants recognized the file, but had forgotten that the file was stored in the cloud. The likelihood a participant remembered a file was stored in the cloud varied significantly based on a number of factors, including the file type, the participant's access to the file (owner, editor, viewer), the file size, and the time since the last modification.

Co-authors Khan and Hyun contributed equally to this work.

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

Participants' responses to our questions about managing files in their cloud storage, as well as those files' sharing settings, revealed a latent need for retrospective data management. Overall, 83% of participants wanted to delete at least one file they saw, while 13% wanted to unshare at least one shared file. To identify features that could help automate retrospective file management in the cloud, we built mixed-effects logistic regression models that attempted to correlate readily available file metadata with participants' decisions, thereby investigating possible predictive factors for these file-management preferences. Beyond a small number of factors (e.g., the participant's access to the file), these straightforward models unfortunately did not capture much of the rationale underlying participants' decisions.

Our study is the first to focus on the need for retrospective file management by grounding questions in a sample of files stored in participants' own cloud-storage accounts. Overall, 81% of participants saw at least one file among the ten in the study that they had forgotten was stored in the cloud, yet responded that it was important to keep that file safe from unauthorized access. Such latent risks are exactly those that users have difficulty effectively understanding or managing. Our study is the first step toward designing interfaces and mechanisms for enabling retrospective file management in the cloud.

RELATED WORK

We summarize cloud storage's history and associated privacy and security concerns. We then describe work to improve retrospective privacy in social media, as well as to manage personal information in email and other archives.

Cloud Storage

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 [35] and ensure data availability, protected from failures [18, 30]. 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. The global market for personal cloud is projected to reach $71.3 billion USD by 2020 [21].

Privacy and Security Concerns for Cloud Storage

Despite its benefits, cloud storage has many implications for privacy and security. A careful analysis of the architecture and workloads of such systems will highlight vulnerabilities in their usage, as well as how these issues impact users [18, 22].

Computer experts have found security issues in the implementation of cloud storage. For example, Hu et al. 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 [26]. 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 [23], 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 [28], the question remains: how do providers implement user-centered data management?

These issues 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 [17]. Although some solutions have been proposed to allow users to take advantage of the cloud without compromising privacy and autonomy [44], 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 [27]. This would explain why users are reluctant to store sensitive data in the cloud [1, 16, 36, 41, 45].

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 toward filling that gap. We investigate cloud storage usage, including why they originally stored files in the cloud, to determine optimal file-management decisions.

Retrospective privacy of social media

While surprisingly little work has investigated retrospective data management for cloud storage, a larger literature has examined analogous questions for social media. Safeguarding privacy in social media is especially complex because users make dynamic privacy decisions based on context [42]. Nevertheless, the social media domain is a useful point of comparison for cloud storage; both support content that can be either shared publicly or kept private.

Temporality mediates whether users perceive content to be public or private. It can also explain the changing relevance of posts over time [2, 49]. The passage of time plays an important role, and users themselves cannot always predict what their preferences will be in the future [7]. Learning from this, we would expect that decisions about sharing files in cloud storage 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 sharing decisions.

In any case, current retrospective mechanisms on social media are limited. Even if users withdraw tweets (e.g., by deleting them), retweets may provide residual evidence and may even highlight when deleted tweets are missing [37]. Cloud storage can create similar problems for users; they may not be fully aware of the consequences of changing file-sharing settings.

User Conceptualization of File Sharing

Local files usually have a single owner who is also the only editor and viewer. In the cloud, however, these privileges can

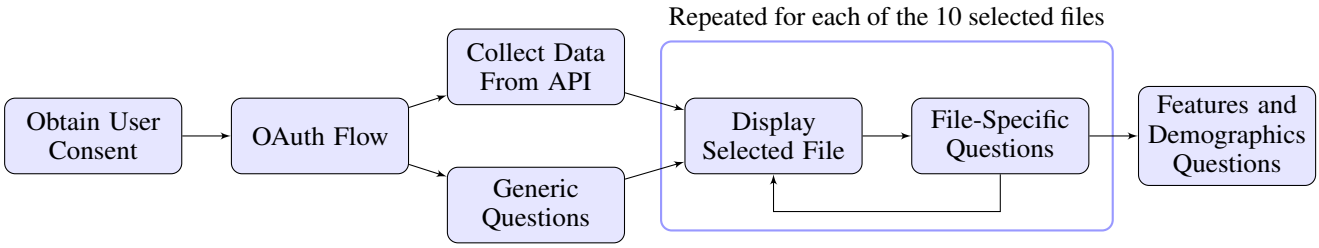

Figure 1: An overview of the survey procedures from the perspective of a participant.

be assigned to others. This distinction is not always intuitive to users, who require more explanation of how file sharing works [15]. Early user experiences with the cloud revealed that users needed to understand a variety of cloud functions, including replication and synchronization, in order to make full use of cloud storage [33].

Nuances of sharing are central to users' understanding of the paradigm of cloud storage [24]. Users often refrain from making decisions about shared-ownership files, even when they can and should, because they relegate authority to the original creator [39, 48]. In particular, it is challenging and confusing for most users to understand the implications of deleting from a shared folder [40]. Keeping copies of files has similar complications because there are two ways to do so: allowing concurrent edits on a single version or reconciling multiple versions [34]. These problems can be exacerbated in shared repositories, and users must develop a variety of management structures and strategies [34].

Personal Information Management

The general research area of personal information management (PIM) began in the 1980s to help users better store, organize, and retrieve collections of data [6, 8, 9, 13, 32]. Researchers have suggested PIM interfaces for web activities [19,29], email [3, 4, 8, 43, 46, 47], and local files [5, 6]. Comparatively little work has focused on PIM for the unique complications of consumer cloud storage.

Critiques of existing PIM systems indicate both technical and usability limitations: they must be adequately supported by current technologies, and users struggle to manage volumes of information constantly increasing over time. Information is compartmentalized between sources, workplaces, and PIM systems themselves [11, 12]. Cloudsweeper, a cloud-based email protection system for PIM, let users remove or "lock up" sensitive, unexpected, and rarely used information. While it effectively protected some sensitive files [43], Cloudsweeper's methods do not map directly to cloud storage. Finally, PIM also raises questions about how users want their files to be managed, especially in group contexts [10, 38].

Determining whether users would want files to be deleted, encrypted, or archived is a complex calculation. Components ranging from file size to access patterns need to be considered to provide efficient and useful file management. Our research contributes to specifying these considerations.

METHODOLOGY

To map out the needs and opportunities for helping users manage forgotten-about files in their cloud storage accounts, our procedure combined programmatic access to the stored files with a dynamic online survey. Due to their popularity and API availability, we chose to implement our survey instrument for both Dropbox and Google Drive. The survey 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 both collected participant demographics and asked about the potential for automating file management. Figure 1 summarizes our survey flow.

Cloud Storage Services

While Dropbox has existed since 2007, Google Drive was introduced in 2012. Both services offer free and paid tiers. Dropbox offers 2GB of free storage, while Google Drive provides 15GB. Google's free 15GB, however, is shared between all Google services, including Gmail and Google Photos.

While the services are similar at a high level, some small differences impacted our study design. 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. How sharing works differs slightly between services: a Dropbox user sharing an individual file can only give others view access; granting edit access requires the entire folder containing the file to be shared. On the other hand, Google Drive allows its users to grant view and edit access for both files and folders. Furthermore, for link sharing, Dropbox users with free accounts are limited to share links with view access only, whereas Google Drive links can apportion view or edit access. When asking specifically about shared files in our survey, we did not consider Dropbox files shared by link because they do not enable collaboration.

Data Collection and Ethics

An essential part of our study involved showing participants files in their own cloud storage accounts, asking questions to 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, which allows our application to programmatically access the files stored within the account. This mechanism allows us to be granted temporary access to these accounts without having to ask users for their passwords. This access can be revoked by the user at any time.

After obtaining authorization, we used the official APIs 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, collecting data in an ethical manner. We 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. When storing the data, we stored only the information we needed, and only stored one-way hashes for any unique identifiers to prevent retaining PII (personally identifiable information). 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.

Recruitment and Inclusion Criteria

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 HIT 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 the participants' accounts were sufficiently well used for us to ask about various use cases.

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.

Index	Selected File Description
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 1: Categories for selecting files in our stratified sample.

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 1. 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.

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. Because Dropbox users cannot share a file for editing via 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.

Survey Structure

Our survey consisted of three main sections. We first asked participants about their usage of cloud storage and general characteristics of their account. 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. This was followed by a set of questions about whether participants recognized the file and, if so, remembered it was in that account. 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 and whether they would want to continue sharing the file with each of them.

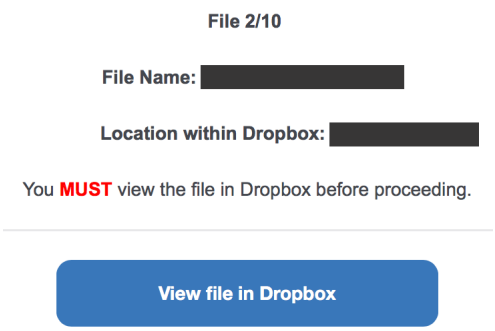

Figure 2: What participants saw at the beginning of each of the ten file-specific sections. Clicking the view button opened a new tab in the browser with a file preview provided by the cloud storage service.

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 participant and, if so, in what circumstances. We have included a more detailed overview of our survey, as well as the full survey instrument, in an appendix in our online supplemental materials.

DATA ANALYSIS

We performed both quantitative and qualitative analyses.

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, 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 (log10) since the file was last modified
- size of the file (log10)
- 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.

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 the body of the paper, we report the p-values for factors that were significant. We provide the full regression tables in the appendix in our online supplementary materials.

RESULTS

We present the results of our survey, as well as our regression models aiming to identify user-specific and file-specific factors that would be predictive of the desired file-management

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

Table 2: Participant demographics

decision and whether the user would remember that file was in cloud storage.

Participants Demographics and Account Usage

Table 2 summarizes the demographics of our participants. 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 3. We also provide distribution plots of these account-level properties in the online supplementary material. While both services have been attracting significant numbers of new users in recent years [25, 31], our participants have been using these services for quite some time; 85% of participants' Google Drive accounts and 94% of their Dropbox accounts were more than 3 years old.

Participants used their accounts in a number of 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 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.

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 files were documents. Documents were far more frequent than spreadsheets and presentations, which accounted for only 3% of 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.

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

Table 3: Descriptive statistics of participants' Dropbox (DB) and Google Drive (GD) accounts.

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 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.

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 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

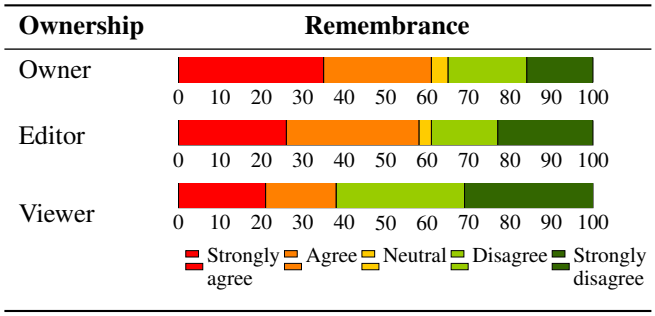

Figure 3: Comparison of file ownership and remembrance (agreement or disagreement that they remembered the file was stored in their cloud account). File ownership had a significant positive correlation with remembering the file was stored in the cloud ($\chi^2(8, N = 862) = 32.244$, $p < .001$).

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 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. Moreover, as detailed in the online supplementary material, we observed a number of significant interactions between file metadata and the cloud-storage service regarding file remembrance.

To investigate the utility of these stored files, we asked participants for a self-reported last accessed time for each file.[1] 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% and 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.

[1]Last access time, as opposed to the time of last modification, is not available via the API.

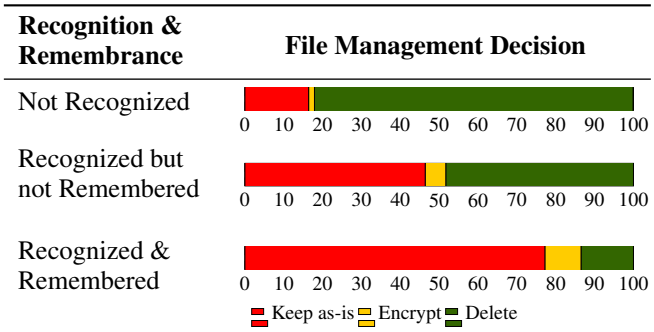

Figure 4: Participants' management decisions across the possible combinations of file recognition and remembrance. These differences are statistically significant ($\chi^2(4, N = 1000) = 260.26$, $p < .001$).

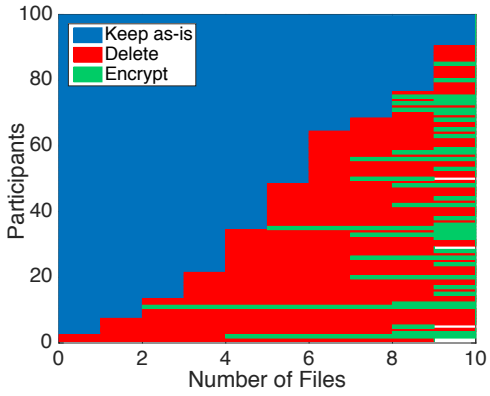

Figure 5: File-management decisions by participant.

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% of 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 is important. Figure 4 demonstrates how these management decisions were significantly correlated with file recognition and remembrance.

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,

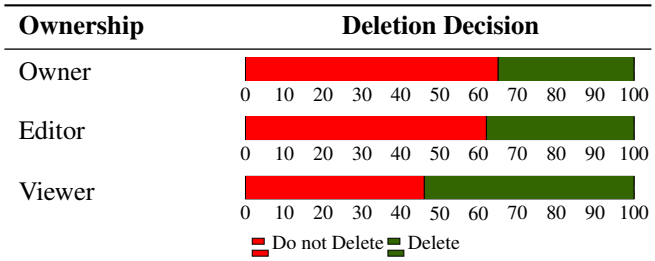

Figure 6: Comparison of deletion and file ownership levels. These values were significantly correlated in our tests ($\chi^2(2, N = 928) = 13.813$, $p < .005$).

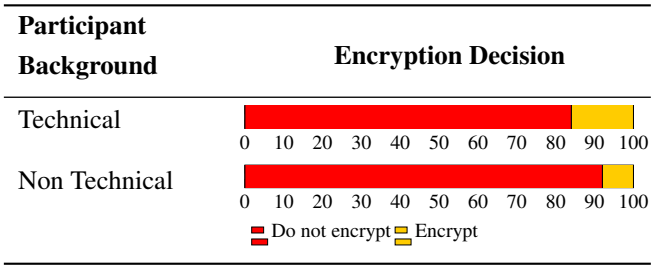

Figure 7: Comparison of of file encryption and the participants' technical background. These values were significantly correlated in our tests ($\chi^2(1, N = 645) = 8.1447$, $p < .005$).

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%). 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 prefer 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. 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. Furthermore, participants who 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.

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 file. The same was true for only 22% of

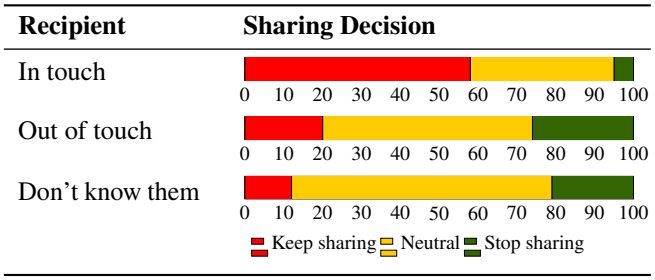

Figure 8: Participants' preferences for definitely continuing to share files (*keep sharing*), not caring whether or not the file continues to be shared (*neutral*), or definitely stopping the file sharing (*stop sharing*) across file-recipient pairs based on whether the participant said they were *in touch* with the recipient (had communicated in the past year), *out of touch* with the recipient (had not communicated in the past year), or did not know the recipient (*don't know them*).

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."

That some files remain 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 partici-

pants 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.

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.

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. Our study highlights why such an automated solution is challenging. 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.

In contrast, participants' free responses explaining how their decisions might generalize suggest that more advanced clustering of files, alongside identifying users' individualized preferences for managing files, might enable partially automated solutions in the future. In particular, our results suggest the need for machine learning approaches that use information extracted from the contents of files to perform more advanced clustering of related files, as well as to identify "useless" files.

Predictive models could combine techniques from machine learning with insights drawn from HCI work on users' security and privacy personas [20]. 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.

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.

CONCLUSION

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). Even so, we found that 83% of participants wanted to permanently delete at least one file from this sample of ten. This result highlights the disconnect between our participants' desired file-management decisions and the high overhead of retrospectively managing thousands of files in a cloud storage account. Thus, our results highlight the need for retrospective privacy mechanisms that empower users to manage the risks latent in their file archives without expending unreasonable effort.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their insightful feedback and Miranda Wei for her assistance. This work was supported in part by the National Science Foundation under grant CNS-1351058.

REFERENCES

1. Ibrahim Arpaci, Kerem Kilicer, and Salih Bardakci. 2015. Effects of security and privacy concerns on educational use of cloud services. *Computers in Human Behavior* 45 (2015), 93–98.

2. Oshrat Ayalon and Eran Toch. 2013. Retrospective privacy: Managing longitudinal privacy in online social networks. In *Proc. SOUPS*.
3. Taiwo Ayodele, Galyna Akmayeva, and Charles A. Shoniregun. 2012. Machine learning approach towards email management. In *Proc. World Congress on Internet Security (WorldCIS)*. 106–109.
4. Olle Balter. 1997. Strategies for organising email messages. In *Proc. HCI*.
5. Deborah Barreau and Bonnie A. Nardi. 1995. Finding and reminding: File organization from the desktop. *ACM SigChi Bulletin* 27, 3 (1995), 39–43.
6. Deborah K. Barreau. 1995. Context as a factor in personal information management systems. *Journal of the American Society for Information Science* 46, 5 (1995), 327.
7. Lujo Bauer, Lorrie Faith Cranor, Saranga Komanduri, Michelle L. Mazurek, Michael K. Reiter, Manya Sleeper, and Blase Ur. 2013. The post anachronism: The temporal dimension of Facebook privacy. In *Proc. WPES*.
8. Victoria Bellotti, Nicolas Ducheneaut, Mark Howard, Ian Smith, and Christine Neuwirth. 2002. Innovation in extremis: Evolving an application for the critical work of email and information management. In *Proc. DIS*.
9. Ofer Bergman, Richard Boardman, Jacek Gwizdka, and William Jones. 2004. Personal information management. In *Proc. CHI Extended Abstracts*.
10. Ofer Bergman, Steve Whittaker, and Noa Falk. 2014. Shared files: The retrieval perspective. *Journal of the Association for Information Science and Technology* 65, 10 (2014), 1949–1963.
11. Richard Boardman and M. Angela Sasse. 2004. Stuff goes into the computer and doesn't come out: A cross-tool study of personal information management. In *Proc. CHI*.
12. Richard Boardman, Robert Spence, and M. Angela Sasse. 2003. Too many hierarchies? The daily struggle for control of the workspace. In *Proc. HCII*.
13. Richard Peter Boardman. 2004. *Improving tool support for personal information management*. Ph.D. Dissertation. University of London.
14. Dell Cameron. 2014. Apple knew of iCloud security hole 6 months before Celebgate. *The Daily Dot*. (September 24 2014). http://www.dailydot.com/technology/apple-icloud-brute-force-attack-march/.
15. Robert Capra, Emily Vardell, and Kathy Brennan. 2014. File synchronization and sharing: User practices and challenges. In *Proc. ASIS&T*.
16. Richard Chow, Philippe Golle, Markus Jakobsson, Elaine Shi, Jessica Staddon, Ryusuke Masuoka, and Jesus Molina. 2009. Controlling data in the cloud: Outsourcing computation without outsourcing control. In *Proc. CCSW*.
17. Jason W. Clark, Peter Snyder, Damon McCoy, and Chris Kanich. 2015. I Saw Images I Didn't Even Know I Had: Understanding User Perceptions of Cloud Storage Privacy. In *Proc. CHI*.
18. Idilio Drago, Marco Mellia, Maurizio M. Munafò, Anna Sperotto, Ramin Sadre, and Aiko Pras. 2012. Inside Dropbox: Understanding personal cloud storage services. In *Proc. IMC*.
19. Susan Dumais, Edward Cutrell, Jonathan J. Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2016. Stuff I've seen: A system for personal information retrieval and re-use. In *ACM SIGIR Forum*, Vol. 49. 28–35.
20. Janna Lynn Dupree, Richard Devries, Daniel M. Berry, and Edward Lank. 2016. Privacy personas: Clustering users via attitudes and behaviors toward security practices. In *Proc. CHI*.
21. Global Industry Analysts. Inc. Accessed 2017. Personal cloud-A global strategic business report. http://www.strategyr.com/MarketResearch/Personal_Cloud_Market_Trends.asp. (Accessed 2017).
22. Glauber Gonçalves, Idilio Drago, Ana Paula Couto Da Silva, Alex Borges Vieira, and Jussara M. Almeida. 2014. Modeling the Dropbox client behavior. In *Proc. ICC*.
23. Graham Cluley. Accessed 2017. Dropbox users leak tax returns, mortgage applications and more. https://www.grahamcluley.com/dropbox-box-leak/. (Accessed 2017).
24. Jane Gruning and Siân Lindley. 2016. Things we own together: Sharing possessions at home. In *Proc. CHI*.
25. Drew Houston and Arash Ferdowsi. 2016. Celebrating half a billion users. https://blogs.dropbox.com/dropbox/2016/03/500-million/. (2016).
26. Wenjin Hu, Tao Yang, and Jeanna N. Matthews. 2010. The good, the bad and the ugly of consumer cloud storage. *ACM SIGOPS Operating Systems Review* 44, 3 (2010), 110–115.
27. Iulia Ion, Niharika Sachdeva, Ponnurangam Kumaraguru, and Srdjan Čapkun. 2011. Home is safer than the cloud!: Privacy concerns for consumer cloud storage. In *Proc. SOUPS*.
28. Eric Johnson. 2017. Lost in the cloud: Cloud storage, privacy, and suggestions for protecting users' data. *Stan. L. Rev.* 69 (2017), 867.
29. Victor Kaptelinin. 2003. UMEA: Translating interaction histories into project contexts. In *Proc. CHI*.
30. Beom Heyn Kim, Wei Huang, and David Lie. 2012. Unity: Secure and durable personal cloud storage. In *Proc. CCSW*.
31. Felix Kollmar. 2017. Cloud Storage Report 2017. https://blog.cloudrail.com/cloud-storage-report-2017/. (2017).

32. Mark W. Lansdale. 1988. The psychology of personal information management. *Applied ergonomics* 19, 1 (1988), 55–66.
33. Cathy Marshall and John C. Tang. 2012. That syncing feeling: Early user experiences with the cloud. In *Proc. DIS*.
34. Charlotte Massey, Thomas Lennig, and Steve Whittaker. 2014. Cloudy forecast: An exploration of the factors underlying shared repository use. In *Proc. CHI*.
35. Peter Mell, Tim Grance, and others. 2011. The NIST definition of cloud computing. (2011).
36. Adriana Mijuskovic and Mexhid Ferati. 2015. User awareness of existing privacy and security risks when storing data in the cloud. In *Proc. ICEL*.
37. Mainack Mondal, Johnnatan Messias, Saptarshi Ghosh, Krishna P. Gummadi, and Aniket Kate. 2017. Longitudinal Privacy Management in Social Media: The Need for Better Controls. *IEEE Internet Computing* 21, 3 (2017), 48–55.
38. Michael Nebeling, Matthias Geel, Oleksiy Syrotkin, and Moira C. Norrie. 2015. MUBox: Multi-user aware personal cloud storage. In *Proc. CHI*.
39. Emilee Rader. 2010. The effect of audience design on labeling, organizing, and finding shared files. In *Proc. CHI*.
40. Kopo Marvin Ramokapane, Awais Rashid, and Jose Such. 2017. "I feel stupid I can't delete...": A study of users' cloud deletion practices and coping strategies. In *Proc. SOUPS*.
41. Esther Schindler. July 2010. Cloud development survey. Evans Data Corporation Strategic Reports. (July 2010).
42. Manya Sleeper, William Melicher, Hana Habib, Lujo Bauer, Lorrie Faith Cranor, and Michelle L Mazurek. 2016. Sharing personal content online: Exploring channel choice and multi-channel behaviors. In *Proc. CHI*.
43. Peter Snyder and Chris Kanich. 2013. Cloudsweeper: Enabling data-centric document management for secure cloud archives. In *Proc. CCSW*.
44. Luke Stark and Matt Tierney. 2014. Lockbox: Mobility, privacy and values in cloud storage. *Ethics and Information Technology* 16, 1 (2014), 1–13.
45. Nabil Ahmed Sultan. 2011. Reaching for the cloud: How SMEs can manage. *International journal of information management* 31, 3 (2011), 272–278.
46. Steve Whittaker, Victoria Bellotti, and Jacek Gwizdka. 2006. Email in personal information management. *Commun. ACM* 49, 1 (2006), 68–73.
47. Steve Whittaker and Candace Sidner. 1996. Email overload: Exploring personal information management of email. In *Proc. CHI*.
48. Hong Zhang and Michael Twidale. 2012. Mine, yours and ours: Using shared folders in personal information management. *Personal Information Management (PIM)* (2012).
49. Xuan Zhao, Niloufar Salehi, Sasha Naranjit, Sara Alwaalan, Stephen Voida, and Dan Cosley. 2013. The many faces of Facebook: Experiencing social media as performance, exhibition, and personal archive. In *Proc. CHI*.