Files of a Feather Flock Together?:
Measuring and Modeling How Users Perceive File Similarity in Cloud Storage

Will Brackenbury
University of Chicago - Computer Science
Masters Thesis
Table of Contents

Chapter 1: Introduction ........................................ 1

Chapter 2: Background and Definitions .......................... 3
  2.1 Perceived Similarity ........................................ 3
  2.2 Co-management ............................................. 3
    2.2.1 Data Similarity ....................................... 4

Chapter 3: Related Work .......................................... 5
  3.1 Similarity .................................................... 5
    3.1.1 Perceived Similarity .................................. 5
    3.1.2 Data Similarity ....................................... 7
  3.2 File Management Tools ...................................... 8
  3.3 Hierarchical Structure ....................................... 10
  3.4 Participant Behavior Classification ........................ 10

Chapter 4: Methodology ........................................... 13
  4.1 Recruitment and Part 1 Survey ............................... 13
  4.2 File Processing ............................................. 13
  4.3 Part 2 ........................................................ 14
  4.4 Analysis Approach ......................................... 15
  4.5 Limitations .................................................. 15
  4.6 Results ...................................................... 16
    4.6.1 Participant Demographics ............................... 16
    4.6.2 Key Characteristics of Participants’ Cloud Accounts .... 16
    4.6.3 Account Organization .................................. 17
    4.6.4 Perceived Similarity .................................. 19
    4.6.5 Additional Perceptions of Similarity .................... 21
    4.6.6 Co-management ........................................ 22

Chapter 5: Construction and Analysis of Classifiers ............ 23
  5.1 Data Similarity ............................................. 23
  5.2 Methodology For Constructing Classifiers .................. 24
  5.3 Predicting Perceived Similarity ............................. 25
  5.4 Predicting Co-management Decisions ........................ 27
  5.5 Data Similarity’s Correlations With Prediction Tasks ....... 28

Chapter 6: Discussion and Conclusion ............................ 31

Chapter 7: Acknowledgements ..................................... 33
List of Figures

4.1 Comparison of file hierarchy structures ........................................... 17
4.2 Perceptions of perceived similarity .................................................. 19
4.3 Perceptions of desired co-management ............................................. 21
5.1 Distribution of data similarity across the file hierarchy ....................... 23
5.2 Summary of other classifier evaluation metrics .................................. 24
5.3 Classification precision, perceived similarity .................................... 25
5.4 Classification precision, co-management .......................................... 27
Abstract

File storage has become effectively free for users, leading to large and disorganized file hierarchies. This scale and disorganization leads to difficulties in retrieving and managing files. One method of alleviating these difficulties is by managing similar files in similar ways (e.g., whether one file being deleted implies the other should be), which we term co-management. Toward an ultimate goal of leveraging this technique to attack these challenges, we conducted a 50-participant online study of how users perceive file similarity and organize similar files in cloud storage. Participants rated the similarity of pairs of files from their Google Drive or Dropbox accounts regarding perceived similarity (e.g., whether the files cover the same topic) and co-management. We correlate perceived similarity and co-management decisions with data similarity (e.g., files’ last-modified dates) and files’ relative locations. We build automated classifiers that identify files similar in perceived similarity or co-management decisions at double or triple the precision of a randomized baseline, and at improved performance against a heuristic baseline. Many of these similar files are located in very different parts of disorganized file hierarchies.
Chapter 1:

Introduction

The cost of digital file storage has continued to significantly decrease. People and enterprises are now able to store more and larger files than ever before. This makes it less feasible for users to manually curate and organize their files, whether on their personal machine [18], in data lakes [14], or in cloud storage [66]. To help with file management tasks, such as file retrieval, researchers have explored methods to predict and highlight directories that might be accessed next by users [26], developed interfaces to quickly access relevant files [61], and investigated the potential of desktop search [7]. These tools, however, have difficulty incorporating contextual information about files, which prior work has identified as vital for file management [6, 4, 13, 14]. Here, we present the first quantitative study to explore the inference and use of such contextual information to organize and manage file storage, through the lens of cloud storage.

We hypothesize that inferring user perceptions of similarity between files can improve the utility of file management tools. For example, if a user located, moved, or deleted a file, they may be interested in co-managing those files: taking the same action on other files they believe to be similar. This notion of similarity may take many forms, such as similarity in topic or purpose. As capturing users’ beliefs of similarity between individual file pairs would be infeasible, a file management tool should infer similarity from both the files’ content and context. Therefore, we further hypothesize that a machine learning classifier could combine various features of data similarity to automatically predict users’ perceived similarity and desired co-management tasks. For example, if there is significant overlap between the lists of shared users for two documents, and their content relates to the same subject, this might imply that the two are a part of the same project and should be co-managed. A critical advantage over prior studies that examined user perceptions of files, is that data similarity may be computed with minimal human intervention based on a range of content- and context-based features: we refer to the numerical scores resulting from these computations as similarity metrics.

To explore the potential of this hypothesis, we conducted a 50-participant online study of Google Drive and Dropbox cloud storage users in which we examined how users perceive the similarity of files within their cloud accounts, how they organize these files, and the extent to which similarity can help with co-management of files. We asked users to rate the similarity of selected file pairs along four dimensions (topic, derivation, purpose, and creation) and compare with how similar the files are based on their content and context. For these files we also asked users to rate the benefit of applying three different co-management actions: finding, moving, and deleting files.
We then trained classifiers to predict these 4 dimensions of perceived similarity and 3 co-management actions. Our results show that, irrespective of the way participants had organized their cloud accounts, our classifiers can predict file similarity and co-management opportunities with precision 1.8x or greater than a random baseline for all seven tasks, and at 2.0x or more for four of the seven tasks. The baseline precision ranged from 14% to 34%, while overall classifier precision ranged from 41% to 67%.

We contribute the following:

- A conceptual framework of file similarity, consisting of Topic, Purpose, Derivation, and Creation Context
- Quantitative analysis of the scale and structure of personal cloud storage accounts
- Qualitative analysis of user perceptions of file similarity
- The creation and evaluation of classifiers for predicting file similarity and co-management

The thesis is structured as follows: In Chapter 2, we provide background and definitions for this work. In Chapter 3, we present prior work in the field, and describe how it compares and differs from what we study here. Specifically, we discuss similarity, file management tools, scale and structure of personal file collections, and classification of user file organization behavior. In Chapter 4, we describe the methodology for the study. We then evaluate survey responses. Specifically, we qualitatively analyze free-responses describing similarity to see how participants conceived of similarity; we analyze the content and structure of participants’ cloud accounts to characterize organization strategies; and we determine the extent to which participants saw files as similar and would like to co-manage files (Chapter 4.6). In Chapter 5, we then construct classifiers, using data features (e.g., filenames, topics expressed in text, and objects identified in images), to identify pairs of files that are perceived to be similar and / or should be co-managed. Finally, in Chapter 6, we conclude by discussing how these findings can aid developers of file management tools, and how this work could be extended.
Chapter 2:

Background and Definitions

We investigate how perceived similarity can be inferred from data similarity in order to address disorganization in cloud storage systems. We define both types of similarity here.

Perceived Similarity

We define perceived similarity to be the various manners in which a user can understand a file as similar or dissimilar. For example, a user may see two documents and recognize that both were created for the same project.

In order to evaluate perceived similarity between files, we draw from various prior works [6, 4, 13, 14] to develop the following framework. It is not exhaustive, but captures many important concepts of perceived similarity from prior literature.

- **Topic:** Two files are similar if they are about the same subject. Example: a photo of a dog and a document about dog grooming techniques.

- **Purpose:** Two files are similar if they will likely be used for similar tasks or purposes. Example: a receipt and a W-2 form saved for tax reporting purposes.

- **Derivation:** Two files are similar if they are different versions of the same item, or if one “created” the other. Example: a paper outline, and the final version of that paper.

- **Creation context:** Two files are similar if they were created at the same time, by the same person, or in the same place. Example: a short story authored at a writer’s retreat, and another person’s poem written at the same retreat.

Co-management

The ultimate goal of this research is to build tools that enable better file management. To that end, we investigate an additional form of similarity: co-management. We define co-management to be the desire to apply of the same file management action to multiple files (e.g. moving two files to the same location). Like perceived similarity, co-management is derived from the users’ perceptions, preferences or intent. We explore three types of co-management:
Chapter 2. Background and Definitions

### Metric Files Description

<table>
<thead>
<tr>
<th>Metric</th>
<th>Files</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Contents</td>
<td>All</td>
<td>Jaccard similarity of chunks of raw file content using MinHash</td>
</tr>
<tr>
<td>Filename</td>
<td>All</td>
<td>Jaccard similarity of the list of bigrams (two-letter chunks) in the filenames</td>
</tr>
<tr>
<td>File Size</td>
<td>All</td>
<td>Logarithm of difference, in bytes, between the file size</td>
</tr>
<tr>
<td>Last Modified</td>
<td>All</td>
<td>Logarithm of difference, in seconds, between the two files’ last modified dates</td>
</tr>
<tr>
<td>Tree Distance</td>
<td>All</td>
<td>The number of steps to reach one file from the other when traversing the file hierarchy (represented as a tree)</td>
</tr>
<tr>
<td>Shared Users</td>
<td>All</td>
<td>Jaccard similarity of the lists of unique user IDs with whom the files have been shared</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text Contents</th>
<th>Text</th>
<th>Cosine similarity between documents’ Word2Vec [48] vector embeddings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Topic</td>
<td>Text</td>
<td>Cosine similarity of documents’ Term Frequency Inverse Document Frequency (TF-IDF) vectors [70]</td>
</tr>
<tr>
<td>Table Schema</td>
<td>Spreadsheets</td>
<td>Jaccard similarity of the column names of spreadsheets, such as .xlsx, .csv, and .tsv files</td>
</tr>
<tr>
<td>Image Contents</td>
<td>Images</td>
<td>Jaccard similarity between unique objects recognized in images by object-detection algorithms [29]</td>
</tr>
<tr>
<td>Image Color</td>
<td>Images</td>
<td>Absolute difference between the average RGB values for each photo</td>
</tr>
</tbody>
</table>

Table 2.1: The data similarity metrics we examined, the files to which they apply, and how we computed them.

- **Find Together**: If a user navigates to one file, they would see other, related, files. This is a form of file retrieval [11].
- **Move Together**: If a user moves one file to another folder, then other, related, files would be moved to the same folder.
- **Delete Together**: If a user deletes a file, then other, related, files would also be deleted.

### Data Similarity

We define *data similarity* to be the set of features that can be extracted from files without human intervention. These features include common file metadata, such as filename and size, as well as content features such as text topics and objects identified in images. We apply various methods to calculate the distance between features such as using the Jaccard similarity between column names in a tabular file. A complete list of the data metrics we considered, and the methods for calculating the distance and similarity values are displayed in Table 2.1. Importantly, these metrics do not require that the user to provide any additional information.
Chapter 3:

Related Work

Our work spans multiple sub-fields. We review the prior literature of each field in the following order: first, we discuss similarity, in the context of both user perceptions and data features. Next, we describe the approaches prior literature has taken to providing file management tools for users. We then detail prior studies investigating the scale and structure of personal file collections—these are relevant to this work, as we conduct a similar investigation on the cloud storage accounts from our user study. Lastly, we describe prior work in classifying user file organization behavior.

Similarity

Perceived Similarity

Prior studies have touched on user perceptions of similarity [63] either in explicit frameworks of user perceptions [40, 4, 6, 14], examinations of file management systems [13, 69, 28], or designs / implementations of file management systems [47, 45, 58].

We attempt to match these frameworks against each other and our own framework of perceived similarity in Table 3.1. We compare items directly against our own framework in the first set of brackets in the table. We also collect several common themes from prior work in the second set of brackets, and note which works discussed these aspects as well. This shows both how our framework builds upon prior work, and which parts of prior work were omitted for this study.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Purpose</th>
<th>Derivation</th>
<th>Creation</th>
<th>Use</th>
<th>Sensitivity</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Perceptions Studies</td>
<td>This study</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Kwasiak, Barreau [40, 4]</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Bergman et al. [6]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Brackenbury et al. [14]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Whitham and Cruickshank [69]</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Golder and Huberman [28]</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>File Management System Studies</td>
<td>Mashwani and Khan [45]</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Mason and Seltzer [47]</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Schroder [58]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.1: Comparison of prior studies’ frameworks of perceptions of context, either through folder organization or user studies. Items in the first set of brackets are from our study, and items in the second set are common themes from other works.
In addition to the dimensions we identified above, we define others from prior work:

- **Use:** Two files are similar if a user would consider the two equally “useful”

- **Sensitivity:** Two files are similar if a user would be equally distressed if the file contents were made public

- **Format:** Two files are similar if they have the same file format or are in the same socially recognized format (e.g. resume, meme).

We included questions about Use and Sensitivity in our user study, but excluded them from our similarity framework, in order to focus on the first four dimensions. We also excluded Format because we felt that it was primarily the domain of data similarity and not perceived similarity.

**Relevant studies**

Kwasnik provided a framework for organizing personal information in [40] by asking 8 academics to give a guided tour of their workspaces. Kwasnik’s study, however, only concerned physical materials. Barreau [4] extended this same framework to electronic materials. This framework contained many more categories and sub-categories than ours, but identified many of the similar attributes that were included in our framework. They identify Topic as “Document Attributes - Topic”, Purpose as “Situation Attributes - Use / Purpose,” and Creation Context as “Situation Attributes - Circumstance.” Most of the additional items that they identified were either data similarity items, such as “Document Attributes - Author” or “Document Attributes - Form,” or they were pieces of contextual information that did not apply to similarity, such as “Situation Attributes - Source.” They did, however, identify Use, Sensitivity, and Format as potential dimensions of similarity.

Boardman and Sasse [13] conducted a longitudinal study of participants’ file organization and management behaviors. They identified several ways participants categorized items into folders: by “project”, “role”, “person”, “topic”, and “document class”. This framework of folder organization works as a framework of perceived similarity. Their concept of “topic” directly matches our framework, and “project” / “person” correspond to Creation Context. “Document class” also corresponds to Format.

Seltzer and Murphy [59] declared the hierarchical file system dead, and laid out several primitives for how new file systems should function. Later, however, Mason and Seltzer [47] acknowledged that this declaration was premature, and described how hierarchical file systems should adopt file relationships as first-class citizens. They describe the relationships of “similar”, “precedes”, “succeeds”, “contains”, “contained by”, and “derived from”. Their relationship of “similar” only refers to data similarity, *data similarity,* another type of similarity, which we introduce later. Otherwise, all the other relationships they introduce are specific instances of Derivation.

Whitham and Cruickshank [69] describe 8 folder functions from their user study. “Exploring groupings and items” matches our description of Topic, “Supporting task execution” resembles Purpose, and “Persistent availability of items” matches Use. The functions, however, are otherwise orthogonal to our framework.
3.1. Similarity

Golder and Huberman [28] describe how tag-based systems capture file relationships akin to our framework. They describe the dimensions of Topic and Purpose as “Identifying who or what it is About” and “Task organizing”, respectively. “Identifying what it is” also resembles Format. The other items in their framework, however, do not relate to similarity or are data similarity.

Mashwani and Khusro [45] describe the implementation of a Semantic File System (SFS), and in the process of doing so, describe concepts that such a file system should support. Among these are items similar to our framework. Most similar to our framework are “tags” (Topic), “temporal usage” and “location-based usage” (Creation Context), and “content similarity” (Derivation). We discuss later how their implementation of these concepts differs from our work presented here.

Schroder et al. [58] demonstrate their implementation of a Semantic Desktop that avoids the cold-start problem (tools have limited utility without user behavior information). They automatically extract concepts from text and use these to build ontologies. These concepts resemble our similarity framework. They describe “project”, “person”, “organization”, “time”, “place”, and “topic”. This naturally aligns with Topic and Creation Context, in our framework, but the other items are not similarity-based.

Bergman et al. [6] describe a user-subjective framework for evaluating information management systems. They describe items as organized by 3 principles: the “subjective classification”, “importance”, and “context” principles. The three, in total, cover our framework. Their framework inspired the development of our own: however, we do not include some items from their framework, such as “social context”, which does not align as well with similarity. Our differing categorization further reflects the ways that similarity differs from subjective attributes about a single file by itself.

In previous work, we [14] devise an explicit framework for file similarity for data lake metadata management. The framework is directly applicable in the context of cloud storage, and it inspires our work here. We do not include here several items from this framework based on the reasons described above.

Beyond these works, there are several threads of work that define context differently. This can either relate to the immediate task that surrounds file management activities [30, 21], or it can refer to the information about the distribution data is drawn from [39, 38]. These works are relevant, but fundamentally different from ours.

Data Similarity

Similarity measures, though often seen in information retrieval and database literature, are used less frequently in connection with each other or to accomplish a higher-level task.

Some prior work describes in general how similarity measures function together. Early work by Jones and Furnas analyzes 7 similarity metrics and their geometric descriptions independent of a larger dataset [37]. Our analysis differs from theirs in that we analyze different measures on a real world dataset. Lee also describes 7 metrics, some of which overlap with those examined by Jones and Furnas [41]. Lee evaluates these metrics on a binary decision task for word co-occurrence. Notably, however, Lee implements this task as a simple decision rule, and does not add more sophisticated modeling techniques. Eck and Waltman additionally examine 14
measures as candidates for normalizing co-occurrence data in scientific literature [23]. They evaluate this on three datasets of co-citations. They also examine the correlation between their similarity measures, as we do in this work. They do not, however, identify this level of similarity with perceived similarity, and again use different similarity measures.

There are also several works that apply similarity measures to prediction and recommendation tasks. The first of these explores semantic similarity between text items via both corpus methods and ontologies [32, 53]. The second explores semantically similar image content via data similarity [55, 49, 56]. The work from Qayyum et al. [55] from this category is particularly relevant: the authors use deep learning methods to retrieve semantically similar images from a medical records database. The primary difference between their work and ours is the restricted nature of their domain, which naturally affects both their effectiveness as well as the relevant features. The others of these focus on a single modality, or work on a functionally different domain, and are not directly comparable.

There are also several works that use data similarity to inform high-level tasks, though they do not necessarily use data similarity as a lens into perceived similarity. Becker et al. [5], for example, use a concert of similarity metrics with clustering methods to identify events in social media streams. Their features match many of ours, but the task is dissimilar. Dong and Halevy use similarity metrics to inform retrieval in a personal information management platform [20]. Notably, however, Dong and Halevy do not examine data from multiple modalities (e.g. text, image) as we do here. Further, the inference techniques they use differ from our own.

**File Management Tools**

The existing literature on file management tools falls into several areas: enhanced navigation, tags, search, and semantic filesystems.

One common evaluation method among these works is measuring how long it takes a participant to retrieve a file in an example task. Because we did not develop a tool, we cannot compare directly on this metric. This will, however, be the evaluation method of choice in our future work.

**Enhanced navigation** Several works attempt to assist navigation of the hierarchical file system by enhancing the interface, often with shortcuts for users to find files they are interested in.

Fitchett and Cockburn [24], for example, describe AccessRank, a tool that takes file system activity as input to identify what files a user will access next. Along with Gutwin [26, 25], they build this into an interface that highlights folders a user is likely to next access. Liu et al. [42] compare against this work and develop BIGFile, a tool that uses similar information to AccessRank, but instead relies on Bayesian information gain as its primary evaluation metric. They additionally create a dual-interface that shows users “shortcuts” to files they might wish to access. Tata et al. [61] build a similar solution into Google Drive, termed QuickAccess, that uses neural networks to make recommendations to the user.

Watanabe et al. [67] proposed FRIDAL, a tool to make file recommendations based on both file access patterns, and also keyword search. They do not, however, look
3.2. File Management Tools

beyond file retrieval. Dumais et al. [22] specifically focus on re-use and re-retrieval with Stuff I’ve Seen, a system they deployed internally to 230 employees. The system provides an interface that integrates multiple information sources, and can provide recommendations for file retrieval based on richer contextual cues (often similar to Creation Context).

All these works are highly related to ours, as they address the same problem, but with different techniques. However, our work also looks beyond retrieval, and uses different features than these tools. These works also use the evaluation method described above, or another that is also not directly comparable to ours.

Bergman et al. [10] investigate and implement a tool, GrayArea, that provides an interface for users to visually reduce the importance of files by placing them in a gray area at the bottom of the interface. They investigate the potential for this solution as an alternative to file deletion. GrayArea, though, does not provide recommendations to users, as we do here.

Tags Prior work has also investigated whether tag-based systems could aid file management. Chou [15] investigates the potential for combining tags with a hierarchical file system in FindFS. Ngo et al. [50] previously described how using tags in this manner could provide another notion of context, and Albadri et al. [2] expand on this work with VennTags, a system for exploring this context using overlapping sets of tags. These different studies, however, do not make direct file recommendations to users. Further, prior work has discovered that users greatly prefer to organize via folders instead of tags, as folders are better at hiding information to avoid cognitive overload, and provide most of needed capability [16, 8]

Search Prior work has also examined whether search capabilities can be used to aid file management. Teevan et al. [62] initially showed that, even with a perfect search capability, many participants still preferred to navigate file hierarchies. Cutrell et al. [17] discuss how contextual cues in Stuff I’ve Seen can enhance search. Bergman et al. [7] conducted a follow-up study 4 years after Teevan’s study, and found that participants still preferred navigation to search. Yet another study by Bergman et al. [9] discovered that this might be because search induces a greater cognitive burden on users. Because many of these works use the earlier discussed evaluation method, we cannot compare results directly.

Semantic Filesystems Work in the semantic web [27] developed into work on the semantic desktop and semantic filesystems [64]. These are systems that incorporate semantic relationships between objects, typically encoded as edges in a graph data structure, into the filesystem. This technique can assist file management techniques, at an increased memory and computational cost.

Mashwani et al. [45] describe the 360 file system, which incorporates information about users’ browsing habits, temporal usage, location-based usage, file movement between directories, file access patterns, content similarity, and tags. Conceptually, this work is highly related to ours. However, relationships such as the content similarity only cover Derivation, and not higher level semantic relationships. Additionally, their implementation only uses heuristics for each individual attribute, and they do not implement any features to combine or control for other metrics.

Jilek et al. [35] demonstrate a similar method by which they create “context spaces,” which encode relationships between similar items along several dimensions. However,
they do not recommend items to users, nor do they provide an evaluation of their system. Many other similar works exist in this area [46, 68], but they display the same differences from our work, and we therefore do not expand on them here.

Hierarchical Structure

Prior studies have examined hierarchical file structure on local storage in the same manner that we do here for cloud storage. We provide an overview in Table 3.2.

Dinneen et al. [18] examined 348 file structures on local storage by using specific software (Cardinal [19]). Zhang and Hu [72] compare the file structures of 2 groups of knowledge workers, administration staff and PhD students (12 participants). Henderson and Srinivasan [33] examine the file structures of 73 knowledge workers using Microsoft Windows in a university setting. Agrawal et al. [1] examined file structure metadata for 5 years of workers at Microsoft, yielding over 60,000 snapshots. Hardof-Jaffe et al. [31] conducted a study of the file structures of 2,081 undergraduate students at Tel-Aviv university via their online storage in the VirtualTAU system. This work is most similar to ours, given that the file storage was on the web. We distinguish ourselves by a more diverse sample population, and the examination of commercial cloud storage systems. We compare directly against the work by Dinneen et al. [18] in Chapter 4.6. We chose this work to compare against because it was the most recent, and contained information about all metrics of interest to our work.

Participant Behavior Classification

A number of prior works have described categorization of users’ file organization habits. We summarize these works in Table 3.3. We focus specifically on studies of digital file collections.

Malone [43] has a categorization that matches ours well: they describe “pilers” and

<table>
<thead>
<tr>
<th># Hierarchies</th>
<th>Study Type</th>
<th>Account types</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study</td>
<td>50</td>
<td>One-time Google Drive / Dropbox accounts</td>
</tr>
<tr>
<td>Dinneen et al. [18]</td>
<td>348</td>
<td>One-time Local storage of diverse sample</td>
</tr>
<tr>
<td>Zhang and Hu [72]</td>
<td>12</td>
<td>One-time Local storage of administration workers and PhD students at a university</td>
</tr>
<tr>
<td>Hardof-Jaffe et al. [31]</td>
<td>2,081</td>
<td>One-time Online storage of undergraduate students</td>
</tr>
<tr>
<td>Henderson and Srinivasan [33]</td>
<td>73</td>
<td>One-time Local storage of Microsoft workers</td>
</tr>
<tr>
<td>Agrawal et al. [1]</td>
<td>&gt; 60,000</td>
<td>One-time Snapshots of local file system metadata from Microsoft employees</td>
</tr>
</tbody>
</table>

Table 3.2: Prior work in file hierarchies

<table>
<thead>
<tr>
<th>More Structured</th>
<th>Less Structured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep</td>
<td>Flat</td>
</tr>
<tr>
<td>Rigid</td>
<td>Fuzzy</td>
</tr>
<tr>
<td>Total filers</td>
<td>Extensive filers</td>
</tr>
<tr>
<td>Occasional filers</td>
<td>Filers</td>
</tr>
<tr>
<td>Structurers</td>
<td>Pilers</td>
</tr>
<tr>
<td>Small folders filing</td>
<td>Big folder filing</td>
</tr>
<tr>
<td>One folder filing</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Prior work in file hierarchies
“filers,” which correspond roughly to flat and deep hierarchies. Boardman and Sasse [13] describe “Occasional Filers,” “Extensive Filers,” and “Total Filers.” Henderson and Srinivasan [33] extend Malone’s framework to also include “Structurers.” Oh [51] classifies participants as “Rigid,” “Flexible,” and “Fuzzy.” Vitale et al. [65] have described participants along a spectrum of “Hoarding” to “Minimalist.” This differs slightly from our classification in that flat hierarchies may still contain many folders, so they do not qualify as truly minimalist.

Hardof et al. [31], however, categorize participant filing behavior at a lower level. They lay out 4 categories: “Piling,” “One-folder filing,” “Small folders filing,” and “Big folder filing.” Roughly, “Piling” and “One-folder filing” correspond to flat hierarchies, and “Small folders filing” as well as “Big folder filing” match deep hierarchies. This is not a perfect match, given that Hardof et al. cluster on average files per folder, largest folder size, root pile rate, and single folder pile rate, whereas we cluster only on number of folders.
Chapter 4:

Methodology

We conducted a two-part online study. The study combined automated analysis of participants’ cloud accounts with surveys heavily incorporating data from the participant’s account.

Recruitment and Part 1 Survey

We recruited participants on Prolific Academic [54], a crowdsourcing marketplace comparable to Amazon Mechanical Turk. We required they be age 18+, live in the USA, and have completed 100+ tasks with 95% approval. We also required participants to have a Google Drive or Dropbox account that was created at least three months prior, contained at least 100 files, and had at least one file shared with someone else.

The study required full read permissions for participants’ cloud accounts, so we took additional measures to ensure they were fully aware of the data we were collecting. Given the known issues with online consent forms [52], we reminded participants multiple times outside the consent form that we were requesting access to analyze files in their account. We also provided an infographic (included in the Appendix) that further detailed our data collection procedures.

Following the consent process, the participant authorized our software to access their cloud account through OAuth. In the short Part 1 survey that followed, we asked general questions about participants’ demographics and usage of their cloud account. This portion took 15 minutes on average. Compensation was $2.50.

File Processing

Once the participant authorized access to their cloud account, we used the Google Drive or Dropbox APIs to analyze the participant’s account, collect file metadata, and compute data similarity metrics between file pairs. If the account contained over 1,000 files, we selected a random, stratified, sample of 1,000 files whose distribution of file types matched that of the full account.

For all files, we downloaded metadata including the last modified date, creation date, and sharing settings. We applied additional extractors to file types of particular interest. For documents (e.g., docx, pdf), we extracted the text. For images (e.g., jpg), we used the Google Vision API [29] to extract the color histogram, recognize
objects in the image and extract any available text. We also treated images with text as documents.

To protect participant confidentiality, we hashed all human-readable information with a participant-specific salt (discarded at the completion of this processing) before writing it to disk. This information included (tokenized) filenames, text extracted from documents, and objects detected in images. This enabled us to identify, for example, that two documents in the account contained the same word or two files were shared with the same user. However, we, as researchers, would not be able to identify that word or user from our collected data.

Once processing was complete, we selected 18 file pairs (set of two files in the participant’s account) to show participants in Part 2. For each of the following criteria, we randomly chose pairs from all files which satisfied the criterion.

- 2 pairs had similar filenames (based on their bigrams)
- 2 pairs’ filenames had a small Levenshtein edit distance
- 2 pairs had a similar set of shared users
- 2 pairs had a similar text topic (based on TF-IDF [70])
- 2 pairs had a similar table schema
- 2 pairs had similar image contents (in Google Vision [29])
- 1 pair was in the same directory (tree distance 0)
- 1 pair was located at tree distance 1
- 4 pairs were selected randomly

Any criterion that did not have the desired number of pairs was replaced with additional random pairs. Similarity thresholds were set via pilot testing.

Part 2

Once we had finished processing the participant’s files, we invited them to complete Part 2. This section was a survey centered on these 18 pairs of files from their own Google Drive or Dropbox account in randomized order. For each pair, we first required that the participant view both files using their cloud accounts’ web-based preview link. The next button was disabled until they opened both links.

We then asked a series of questions about that pair. First, we asked the participant to describe each file in free text. We then asked them to describe in free text how they believed the files were similar or dissimilar. Next, we asked them to rate their agreement with a series of statements on five-point Likert scales (“strongly agree” to “strongly disagree” with a sixth “don’t know” option). This series included statements about our four classes of perceived similarity (e.g., “I consider these two files to be similar in Topic”). It also included statements about our three types of co-management (e.g., “If I were searching for information, and I found one of these files to be relevant, I would also want to see the other file”).

Part 2 took about one hour to complete. We compensated $10.00.
4.4. Analysis Approach

Analysis Approach

In part 2 of our study, we collected both quantitative and qualitative data. In addition, we collected data similarity for many additional file pairs in part 1. Some of our quantitative analysis is purely descriptive (e.g., summarizing how accounts are organized or how files perceived as similar are distributed across the file hierarchy). As appropriate, we perform statistical testing (described more fully in our results).

Our construction and evaluation of machine learning classifiers to predict files’ perceived similarity or desired co-management is another key quantitative analysis. We present our detailed methodology for this task later in the thesis.

We analyzed our free-text response data primarily through qualitative coding. To surface participants’ general conceptions of file similarity, for each file pair we asked participants to “please describe in general how these files are similar or dissimilar.” While we developed our four classes of perceived similarity based on the literature [14, 13, 4, 6], we did not have concrete hypotheses about other types of similarity that might emerge from the data. Therefore, two members of the research team conducted collaborative affinity diagramming [12] of free-text responses. The two coders iteratively clustered and re-clustered the 900 responses until both were satisfied each cluster represented a distinct theme. This process resulted in 15 initial clusters. The two coders then collaboratively created sub-clusters within each initial cluster to further disentangle a theme’s nuances. This process resulted in 46 sub-clusters.

Limitations

We report on a convenience sample of crowdworkers, which is not representative of any broader population. Despite our efforts to clearly communicate how our data collection respected the privacy of participants’ cloud accounts, privacy-conscious crowdworkers were likely reluctant to take part in our study, further biasing our sample. Because we asked the same questions for each file pair, participants may have been prone to fatigue and inattention. We worked to mitigate this concern by iteratively shortening both multiple-choice and free-response sections of the survey throughout extensive pilot testing.

Because most randomly selected pairs of files in a cloud account are unlikely to be similar, we used stratified sampling, rather than fully random sampling, to select file pairs. As a result, our quantitative analyses likely overestimate the likelihood two random files in a cloud account are similar. However, this oversampling enabled us to investigate similarity more deeply with a fixed sample size. Across participants, we collected data similarity for over 11 million file pairs and both perceived similarity and co-management responses for a total of 900 file pairs. While large for a user study, this represents relatively small data for machine learning. Collecting more data is likely to improve our classifiers’ performance.
Chapter 4. Methodology

### Results

We first present participant demographics and describe key characteristics of participants’ cloud accounts. We then explore how participants organized files within their accounts. This includes both self-reported perceptions, as well as quantitative analysis of accounts’ folder hierarchies. Finally, we explore participants’ quantitative ratings and qualitative explanations of both perceived similarity and desired co-management.

#### Participant Demographics

In total, 50 participants completed our full protocol. Of the 50, 27 (54.0%) identified as female, 20 (40.0%) identified as male, and 3 (6.0%) identified as non-binary or other. The most common age range was 25–34 years old (48%) and the second most common range was 18–24 years old (30%). 13 (26.0%) had held a job or taken a course in computer science.

Among the participants, 46 (92.0%) used Google Drive accounts for the study. The other 4 used Dropbox. Regarding typical account access, 98.0% of participants used the service’s web app, 60.0% used the mobile app, and 30.0% had automatic sync active. 17 (34.0%) were daily users of their primary service, 25 (50.0%) were weekly users, 7 (14.0%) were monthly users, and 1 chose not to respond. On average, participants estimated that personal data was 74.0% of their account, while 25.4% was professional data.

#### Key Characteristics of Participants’ Cloud Accounts

Participants’ accounts were used for various purposes and contained a wide range of content. Table 4.1 reports general characteristics. Our 50 participants had a total of 119,388 files in their accounts. We observed 341 unique file extensions. We grouped file extensions for analysis, as shown in Table 4.1. The most common file type was images (72,125 files), mostly split between jpg (46,019) and png (22,422) files. Second was a catch-all “other” category (14,729). The most common “other” file extension

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>25th Percentile</th>
<th>Median</th>
<th>75th Percentile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account Age (days)</td>
<td>148.1</td>
<td>2,404.69</td>
<td>3,000.59</td>
<td>3,570.09</td>
<td>4,546.44</td>
</tr>
<tr>
<td>Total Size (Gb)</td>
<td>0.39</td>
<td>1.69</td>
<td>4.67</td>
<td>11.14</td>
<td>151.31</td>
</tr>
<tr>
<td># Files</td>
<td>123.00</td>
<td>297.50</td>
<td>540.50</td>
<td>1,444.75</td>
<td>17,081.00</td>
</tr>
<tr>
<td>(# images)</td>
<td>0.00</td>
<td>32.75</td>
<td>168.00</td>
<td>731.75</td>
<td>15,123.00</td>
</tr>
<tr>
<td>(# documents)</td>
<td>4.00</td>
<td>44.00</td>
<td>139.50</td>
<td>298.0</td>
<td>2,345.00</td>
</tr>
<tr>
<td>(# spreadsheets)</td>
<td>0.00</td>
<td>5.00</td>
<td>15.50</td>
<td>33.75</td>
<td>207.00</td>
</tr>
<tr>
<td>(# presentations)</td>
<td>0.00</td>
<td>0.00</td>
<td>2.50</td>
<td>10.0</td>
<td>152.00</td>
</tr>
<tr>
<td>(# web files)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.75</td>
<td>2,453.00</td>
</tr>
<tr>
<td>(# media files)</td>
<td>0.00</td>
<td>6.25</td>
<td>40.50</td>
<td>129.25</td>
<td>5,532.00</td>
</tr>
<tr>
<td>(# other files)</td>
<td>0.00</td>
<td>7.25</td>
<td>25.00</td>
<td>87.0</td>
<td>10,060.00</td>
</tr>
<tr>
<td># Folders</td>
<td>3.00</td>
<td>9.25</td>
<td>27.00</td>
<td>95.0</td>
<td>3,185.00</td>
</tr>
<tr>
<td># File extensions</td>
<td>5.00</td>
<td>12.00</td>
<td>15.00</td>
<td>24.00</td>
<td>75.00</td>
</tr>
<tr>
<td># Users shared with</td>
<td>2.00</td>
<td>267.00</td>
<td>627.00</td>
<td>4,056.50</td>
<td>215,852.00</td>
</tr>
</tbody>
</table>

Table 4.1: Key characteristics of participants’ cloud accounts.
4.6. Results

(a) Flat hierarchies.  
(b) Deep hierarchies.

Figure 4.1: Visualization of the typical folder structure of flat and deep hierarchies. These trees merge participants’ folder structures and show commonalities contained in all accounts (dark blue nodes) through 20% of accounts (white nodes).

was flat (3,664), which is a database file format. Documents (13,405) and media files (12,334) followed. The median account age was over 8 years, and the median account size was over 4 Gb. Together, this suggests that these accounts were actively used for diverse purposes over many years.

Participants’ free-responses describing files shown during the study underscored the diversity of account uses. Participants reported personal items like wedding photos (“my brother, cousin and uncle in Central Park after our wedding ceremony”) and entertainment. They had professional products (“an article I wrote”) and job-seeking materials (“information...for applying for a job”). Diversity was high even within a given account. For example, one mixed “a practice exam for Organic Chemistry 1” and “an mp3 of the song Come With Me Now,” while another mixed “my timesheet to get paid” and “a list of aviation analysts who have given comments to the press about Boeing crashes.”

Account Organization

Although alternatives like keyword search serve important purposes in personal information management [59], hierarchical file systems remain the default way of organizing files [47]. We therefore both examined participants’ perceptions of the organization of their account and measured their accounts’ actual hierarchies. These analyses help us contextualize the distribution of similar files throughout participants’ accounts.

Participants were split on whether they considered their account well-organized. While 36% agreed with a statement that their account is well-organized, 40% disagreed and 24% were neutral. Participants frequently described careful usage of folders to justify considering their account well-organized (e.g., “I name the folder of the topic what the photos or files fall under. Then when it gets too full/I can’t easily get to new files, I create a subfolder to put related content in”). They also reported strategies like organizing files by date. In contrast, some participants who felt their account was not well-organized described intentionally choosing not to use folders (e.g., “With text search and picture view I find it irrelevant”). Others reported failed attempts at using folders (e.g., “I have folders I use to split up files, but I threw everything up
there when my laptop was failing and I have to go back and reorganize it”).

These free-response answers suggested a dichotomy in participants’ use of hierarchical folders to organize accounts. Therefore, we measured how they structured folders in their actual cloud accounts. We performed k-means clustering on the number of folders per account, finding a threshold of 10 folders. We term accounts containing 10 or fewer folders flat as their lack of folders implies keeping most files in one place. In contrast, we term accounts with over 10 folders deep.

Among participants, 16 (32.0%) had flat hierarchies, while 34 (68.0%) had deep hierarchies. Figure 4.1 visualizes the typical folder hierarchy for both classes. We created this figure by converting each account’s folder structure to a tree and then merging these into a single tree greedily beginning with the deepest branch (the one with the longest chain of sub-folders). Each node in this merged tree represents a folder. We colored each node based on the proportion of accounts that contained a folder in that position, pruning all nodes that appeared in under 20% of accounts. As shown in Figure 4.1, flat hierarchies typically contained the root directory and one or two sub-folders. In contrast, most deep hierarchies contained many sub-folders off the root directory, plus a few deeper branches.

Table 4.2 further quantifies differences between flat and deep hierarchies. Similar to prior work on file systems [18], using the Shapiro-Wilk test we found that many metrics of account structure were lognormally distributed across participants. For these, we report the adjusted mean \( e^\mu \), standard deviation \( e^\sigma \), and median \( e^{med} \) of the log-distribution. Unsurprisingly, the average flat hierarchy has many more files per folder than the average deep hierarchy. In deep hierarchies, files in the same directory are more likely to have the same file extension.

Interestingly, 15 of the 18 participants (83.3%) who agreed that their account was well-organized had a deep hierarchy. In contrast, participants with both flat hierarchies and deep hierarchies did not agree that their account was well-organized.

As our goal is to understand perceived similarity and co-management relative to both data similarity and files’ locations in the folder hierarchy, the distinction between flat and deep hierarchies is important. Trying to locate a file in a flat hierarchy is different than in a deep one. Furthermore, trying to find a file nearby in a folder hierarchy is likely easier than trying to find one far away, meaning that tools that aid in the latter can be particularly valuable. Thus, in the remainder of our analyses,

<table>
<thead>
<tr>
<th></th>
<th>Deep</th>
<th>Flat</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e^\mu ) # files</td>
<td>1,146.00</td>
<td>311.00</td>
</tr>
<tr>
<td>( e^\mu ) # folders</td>
<td>75.00</td>
<td>6.00</td>
</tr>
<tr>
<td>Mean files per folder</td>
<td>25.00</td>
<td>69.00</td>
</tr>
<tr>
<td>Mean depth</td>
<td>3.07</td>
<td>1.01</td>
</tr>
<tr>
<td>Mean breadth</td>
<td>9.13</td>
<td>2.37</td>
</tr>
<tr>
<td>( e^\mu ) # unique file extensions</td>
<td>26.00</td>
<td>11.00</td>
</tr>
<tr>
<td>( e^\mu ) # unique folders per extension</td>
<td>8.44</td>
<td>1.32</td>
</tr>
<tr>
<td>( e^\mu ) # unique extensions per folder</td>
<td>1.86</td>
<td>4.13</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison of account characteristics by hierarchy. Depth is the number of clicks needed to reach a file from the root. Breadth is the number of subfolders in a folder.
4.6. Results

Table 4.3: Comparison of the lognormally distributed (top) and normally distributed (bottom) characteristics of local storage (taken from Dinneen et al. [18]) vs. cloud storage (our results).

<table>
<thead>
<tr>
<th></th>
<th>$e_{med}$</th>
<th>$e_{\mu}$</th>
<th>$e_{\sigma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local</td>
<td>Cloud</td>
<td>Local</td>
</tr>
<tr>
<td># files</td>
<td>29,123</td>
<td>540</td>
<td>193,001</td>
</tr>
<tr>
<td>Max breadth</td>
<td>947</td>
<td>12</td>
<td>4,990</td>
</tr>
<tr>
<td>Mean breadth</td>
<td>290</td>
<td>4</td>
<td>888</td>
</tr>
<tr>
<td>Leaves</td>
<td>2,582</td>
<td>20</td>
<td>18,192</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local</td>
<td>Cloud</td>
</tr>
<tr>
<td>Max depth</td>
<td>15.45</td>
<td>5.36</td>
</tr>
<tr>
<td>Waist depth</td>
<td>6.27</td>
<td>2.18</td>
</tr>
<tr>
<td>Mean leaf depth</td>
<td>7.00</td>
<td>2.66</td>
</tr>
<tr>
<td>Percentage leaves</td>
<td>73%</td>
<td>70%</td>
</tr>
</tbody>
</table>

Figure 4.2: Participants’ agreement that the file pairs they saw exhibited the four types of perceived similarity, broken down by whether the file pair was chosen due to a similarity metric exceeding a threshold, or randomly.

we distinguish between the 252 file pairs in flat hierarchies (flat), the 227 in nearby directories in deep hierarchies (deep:close), and the 421 not in nearby directories in deep hierarchies (deep:far). We classify file pairs with tree distance at most 2 as deep:close, and the others as deep:far.

While prior work on cloud accounts is relatively limited, a larger literature has studied how local file systems (e.g., a laptop’s hard drive) are organized. Table 4.3 compares key characteristics of participants’ cloud accounts with findings from prior work on local file systems [18]. Cloud accounts were much smaller in size, folder breadth, and folder depth than local storage, aligning with prior characterizations of differences between cloud and local storage [34].

Perceived Similarity

We now explore participants’ perceived similarity between file pairs. We first examine each of the four similarity types suggested from the literature. For these, we collected both quantitative (Likert-scale responses) and qualitative data. Figure 4.2 displays a breakdown of the former by the method of file pair selection. We see that participants more often responded “Strongly agree” or “Agree” when the file pair was selected via a similarity metric. We also see that many more pairs were perceived as dissimilar than were perceived as similar.

The qualitative data let us unpack nuances of that type of perceived similarity.
Chapter 4. Methodology

Perceived Similarity

<table>
<thead>
<tr>
<th></th>
<th>Topic</th>
<th>Derivation</th>
<th>Purpose</th>
<th>Creation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>—</td>
<td>0.387</td>
<td>0.594</td>
<td>0.479</td>
</tr>
<tr>
<td>Derivation</td>
<td>0.387</td>
<td>—</td>
<td>0.359</td>
<td>0.510</td>
</tr>
<tr>
<td>Purpose</td>
<td>0.594</td>
<td>0.359</td>
<td>—</td>
<td>0.555</td>
</tr>
<tr>
<td>Creation</td>
<td>0.479</td>
<td>0.510</td>
<td>0.555</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 4.4: Pearson correlation of perceived similarity responses.

We additionally perform a Pearson correlation test between each type of perceived similarity to find interrelationships—we report this data in Table 4.4. Each test was significant at the p<0.001 level. We then discuss additional types of perceived similarity that emerged in our qualitative analyses.

**Topic similarity** Participants labeled 31.7% of file pairs they saw as similar in topic. Topic similarity was most strongly correlated with purpose similarity. For one file pair, a participant said they “are similar because they are both related to aquariums and I sold both of them to aquarium blogs.” Another said they were similar as “they are both original articles I wrote to inform people about CBD.” Participants extended the idea of topic similarity beyond a conceptual subject, like aquariums or CBD, to the subjects appearing in photos. For example, a participant said two photos were similar “because my son is in both files and with my husband in one and myself in the other.” Another said, “These are very similar, both are lawn equipment that I am attempting to sell.” However, participants sometimes described how subjects of photos differed based on other factors: “These both contain the same character from the same anime, so they’re similar in topic. One exists within the universe of the show and the other is more based on a fan’s vision…”

**Purpose similarity** Participants indicated 36.9% of file pairs they saw were similar in purpose. Purpose similarity was most strongly correlated with topic similarity. Many purposes described were very concrete, such as “they are both preproduction documents for a short film.” Files could be similar in purpose despite being different otherwise. For example, one participant said two files are “very different besides that they both are for entertainment.” Another stated, “Both serve a similar purpose, to keep track of something, but topics are completely different.” Participants invoked similarity in purpose for both concrete tasks (e.g., “they are both pictures of Arnold that were used to troll political figures”) and conceptual goals (e.g., “both related to fitness and making a workout plan”).

**Derivation similarity** Participants identified 13.8% of file pairs they saw as similar in derivation. Derivation was most strongly correlated with similarity in creation context. They described derivation similarity in many ways. One was that a file was an extension of another. A participant responded, “The first file is part of the full set that the latter contains, they are from the same project for the same class.” Another was that files contained overlapping items, but were different instances. For example, “They are both match recordings of me at the same event. I also lose both games in the footage.” Another conceptualization was that the files referred to very similar items used for the same task. For example, “They are photos of the same dollar store I took at roughly the same time at the same angle.” Lastly, participants
labeled many files that are duplicates or minor variations of each other as similar in derivation.

**Creation context similarity** Participants reported 24.1% of the file pairs they saw as similar in creation context. Derivation is most strongly correlated with creation type similarity. For example, “These files are extremely similar, both about tarot cards and both written for work and the same client.” Sometimes, many aspect of the creation context aligned (e.g., “These are both photos of me with my dog, taken on the same day and in the same location”). In other cases, the alignment was only partial (e.g., “One’s in korean the other english. One has people, the other is just text. Both are from a korean comic”). When and where aspects of the file originated was crucial. For example, a participant reported a file pair as dissimilar for the following reason: “They are both icons of fanart of the same character. The only difference is that they are by different artists.”

**Additional Perceptions of Similarity**

Our qualitative analysis surfaced additional types of perceived similarity beyond those we had predicted for our framework. First, participants reported similarity in presentation for files depicting their subject in similar manners. We observed variants based on format, as well as on language and style. Similarity in format was sometimes based simply on file type (e.g., both being photos), but sometimes more nuanced (“both are basically reminders in picture form”). Similarity in language and style encompassed writing style and tone, as well as genre. In an example of dissimilarity, a participant wrote, “They are similar in that they are both music. They are dissimilar in that they are both different genres.” Another dimension of similarity was how “complete” the file was. One participant described a file pair as “both Asian. both from comics. one black and white one in color.”

Other responses highlighted similarities in the implications of files in a pair, including the sensitivity of the file. For example, one participant wrote, “It could be said they are similar because of an intimate context.” The participant described the files as a photograph of two hands making contact and an e-book about love. Additionally, files were often categorized as similar in sensitivity if they were both public or both private. For example, “One is a spreadsheet and the other is a doc. One is private and the chat was shared.”
**Co-management**

We examine Likert responses for co-management in Figure 4.3. We see similar trends as for the quantitative data for perceived similarity (Figure 4.2).

Participants expressed the desire to co-manage files for 21.8% to 32.6% of all file pairs shown, depending on the management action. Further, the instances in which they wished to make a co-management action were highly correlated with other co-management actions, as well as with their perceived similarity of the file pair. We investigated this with Pearson correlation tests between each pair of co-management actions, and between every action and perceived similarity type. The coefficients are reported in Table 4.5. Every coefficient was significant at \( p < 0.001 \) post correction for multiple discovery.

**Find together:** For 25.4% of the file pairs they saw, participants agreed that if they were accessing one file in the pair, they would also want to access the other. Finding files correlated weakly with moving files. This action would allow users to more easily retrieve files likely to be related. One participant described this by stating, “These files are extremely similar as they are both written for work and supplied to the same agency to be given to clients.”

**Move together:** For 32.6% of the file pairs they saw, participants agreed that if they were moving one to a different folder, they would also want to move the other. Moving files together was most strongly correlated with creation context. This action could equip users to better organize files conceptually using the folder hierarchy. One participant stated that two files drawn from the same folder “were put together at the same time as part of a SWOT exercise when choosing a wedding location.”

**Delete together:** For 21.8% of the file pairs they saw, participants agreed that if they were deleting one, they would want to delete the other. However, there was little correlation between perceived similarity and deletion. We posit that this is because deletion is a more permanent operation that the others.

This action might enable users to clean up their repository more efficiently, or to better remove all traces of sensitive files they might want to delete.

<table>
<thead>
<tr>
<th>Co-management</th>
<th>Perceived Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find</td>
<td>Move</td>
</tr>
<tr>
<td>Find</td>
<td>—</td>
</tr>
<tr>
<td>Move</td>
<td>0.685</td>
</tr>
<tr>
<td>Delete</td>
<td>0.533</td>
</tr>
</tbody>
</table>

Table 4.5: Pearson correlation coefficients within co-management preferences, as well as correlating co-management with perceived similarity.
Chapter 5:

Construction and Analysis of Classifiers

We explore in this section the use of data similarity to predict perceived similarity and file co-management. We find that many data similarity metrics are not correlated with tree distance, implying that files with high data similarity exist in different parts of the hierarchies. We also find that a classifier using several data similarity metrics as features can predict perceived similarity and file co-management in seven representative tasks. To better understand the source of our classifiers’ predictive power, we end this section by constructing logistic regression models for each of the seven prediction tasks.

Data Similarity

We collected 11,653,450 pairs of similarity metrics for all file types, with an additional 4,519,675 pairs for image similarity metrics, 4,262,444 for text similarity metrics, and 39,333 for the table similarity metric.

We examine the distribution of data similarity against tree distance to gauge how it varies throughout the hierarchy in Figure 5.1. All but three metrics are naturally distributed between 0 and 1, which we denote as a similarity value. Three metrics (File Size, Last Modified, Image Color) are instead a distance metric. We therefore convert these to a similarity value by normalizing each against its maximum observed value across all participants, and subtract this score from 1.

There are two key takeaways from Figure 5.1. First, for most metrics, they are more often dissimilar than they are similar. File Content, in particular, has very few files that are similar, and its median value across all tree distances is very small.

![Figure 5.1: Box plots depicting how each class of data similarity is distributed for all file pairs in participants’ accounts. The box plot labeled F shows the distribution for all pairs in flat accounts. The remaining box plots represent the distribution in deep accounts at the tree distance specified by the label (e.g., “0” represents the distribution for file pairs in the same directory.)](image-url)
Second, none of the metrics are strongly correlated with tree distance. This implies that files with high data similarity exist in different parts of the hierarchies.

Methodology For Constructing Classifiers

We develop seven classifiers that predict four types of perceived similarity and three instances of file co-management. We also identify metrics for evaluation that demonstrate these techniques’ potential to provide file co-management recommendations.

Our seven target variables are encoded as 5-point Likert data from “Strongly disagree” to “Strongly agree,” with an “I don’t know” option. We recode “Strongly disagree” / “Disagree” / “Neutral” as 0, and “Agree” / “Strongly agree” as 1. We drop “I don’t know” responses. This coding is done to represent how a tool developed in the future might make recommendations. Either a file pair is similar, and the tool should emit the pair to the user, or it is neutral / not similar, and should be ignored. Whenever we subsequently state that a participant believed a file pair had a particular type of perceived similarity, we mean that their response to the associated survey question was “Strongly agree” or “Agree.”

We evaluate our classifiers using several metrics:

1. **Precision** We evaluate as the mean of precision scores for all participants.

2. **Number of pairs successfully predicted to be similar** We examine the number of file pairs that the classifier successfully predicts to be similar. We compare against the number of file pairs that each participant labeled as similar (Table 5.2a).

3. **Tree distance of pairs successfully predicted to be similar** We take the tree distance of file pairs where both the classifier predicted the pair to be similar and the participant stated that pair was similar. If this tree distance is not small (i.e. > 2), then the classifier is able to predict similar file pairs that are far apart in the hierarchy (Table 5.2b).

We further average all these scores over five instances of the same classifier trained via 5-fold cross validation. We then report these metrics for the subgroups of “flat,” “deep:close,” and “deep:far.”
5.3. Predicting Perceived Similarity

Figure 5.3: Classification precision in predicting participants’ \textit{perceived similarity} of file pairs from those pairs’ data similarity.

To evaluate the \textbf{Precision} of each classifier, we compare against a randomized baseline, and a heuristic baseline. The randomized baseline is the precision of a classifier that uniformly randomly assigns the label 1 to file pairs from the test set. The precision of this is equal to the probability that a randomly chosen file pair will be labeled as correct by the participant. The heuristic baseline assigns the label 1 to all pairs where at least half the data similarity metrics have a value above 10% of the maximum observed value for that metric. Intuitively, beating this classifier represents learning a signal about feature importance, and feature interaction. We also focus on precision, rather than accuracy or recall because we aim to provide a few, high quality recommendations and we lack a large labeled dataset.

The large class imbalance for these seven tasks, where many more file pairs are dissimilar than similar, leads to a random baseline with lower precision. We note that our labeled dataset consists of file pairs specifically selected so as to have perceived similarity. This implies that the class imbalance of the underlying data would be much larger. Because our classifier improves with more data, our evaluation of precision likely underestimates the true precision that would be achieved on a larger labeled dataset.

We evaluate four classifiers for these target tasks: linear-kernel SVM, RBF-kernel SVM, random forests, and boosted trees. Because boosted trees had the highest overall precision for both predicting perceived similarity and co-management, we report results only for boosted trees (termed \textit{the classifier}) in the remainder of the paper.

\section*{Predicting Perceived Similarity}

Figure 5.3 shows the results of our classifier compared with the random baseline and heuristic baseline for different organizations (i.e., deep or flat) and distances (i.e., close vs far). Overall, we exceed the random baseline by 1.8x or more for all seven tasks, and we exceed it by 2.0x or more for four of the seven. However, we only exceed the heuristic baseline in four of the seven tasks, and by much slimmer margins. The randomized baseline precision ranged from 14\% to 34\%, and the heuristic baseline precision ranged from to 37\% to 77\%. Overall classifier precision ranged from 41\% to 67\%.

There is no clear winner among flat, deep:close and deep:far hierarchies as to which yields the highest precision scores. This is a promising result: it indicates that the tool is effective across hierarchy types, and that it is learning signal beyond any that
might be present in the organizational structure. Because the predicted distribution of “number of predicted items” and “tree distance of successfully predicted file pairs” were similar to the underlying labeled data, one can further state that the classifier was successful on diverse subgroups of data. We discuss each of these results below. In most cases, these metrics match the mean and standard deviation of participant labels. Further, the results based on tree distance demonstrate that the classifier is able to find some items that are far apart in the file hierarchy. This greatly enhances the potential utility of this technique. It is also notable that the mean predicted tree distance for the heuristic classifier was significantly lower than the equivalent metric for the classifier and the underlying data. This indicates that, even if the heuristic baseline might have higher precision, the recommendations it would provide are far less useful. Additionally, the heuristic baseline recommends significantly fewer file pairs on average than the classifier or random baseline, which further limits the heuristic’s utility.

Figure 5.3a shows results for topic similarity. The overall classifier had an average precision of 63.5%, beating the random baseline by 2.0x. However, the heuristic baseline had the highest precision with 70.6%. The mean number of file pairs the classifier successfully predicted as similar was 2.88 (sd 2.19). This is lower than the mean number of items participants labeled (5.84), but it exceeds the mean of 2.10 from the heuristic classifier. The mean overall tree distance of successfully predicted file pairs was 3.33 (sd 2.58). This is nearly identical to the mean tree distance between file pairs labeled by participants which was 3.68 (2.88). The mean overall tree distance for the heuristic baseline was 2.45 (1.36), which was significantly lower. The minimum for both baselines and the classifier was 0, and the maximum for the classifier, the random baseline, and the heuristic were 21, 14, and 6, respectively. Notably, the classifier outperforms the random baseline for deep accounts with close and far file pairs by 4.7x and 2.7x, whereas it only outperforms the random baseline by 1.5x for flat accounts. Similarly, the classifier does not exceed the heuristic baseline for flat and deep:close pairs, but does exceed it for deep:far pairs. This discrepancy suggests that folders may often be used to organize items by topic similarity.

Figure 5.3b shows the results for purpose similarity. The classifier had an average precision of 66.9%, which does not exceed the heuristic baseline at 76.7%. The mean number of successful similarity predictions it made was 3.83 (sd 2.16, max 9), which compares favorably to the heuristic baseline at 1.93 (sd 0.74, max 5). Both of these are lower than the number of items labeled by participants, indicating the potential for optimization. The mean tree distance of successfully predicted file pairs was 3.57 (max 21), which again compares favorably to the heuristic baseline at 2.78 (max 8). The results are highly promising for both flat accounts, as well as deep:far file pairs, beating random baselines by 2.5x and 2.6x, respectively, but the most interesting result is for deep:close pairs, whose baseline we beat by 4.6x. The mean precision for deep:close pairs is 80%. This is particularly impressive given that the random baseline is lower than for many other types of perceived similarity. However, these results are colored by the fact that the heuristic baseline exceeds or matches nearly all of these results. This indicates that the classifier has learned a signal that is independent of the file hierarchy, but not one that requires complicated feature interaction.

Figure 5.3c shows results for derivation similarity. While the classifier outperforms
5.4. Predicting Co-management Decisions

We now turn our attention to predicting co-management decisions. Figure 5.4 shows the results of the classifier versus the random baseline and heuristic baseline for different hierarchies (i.e., deep or flat) and distances (i.e., close vs far). Our classifier outperforms the random baseline for all co-management decisions and hierarchies, and outperforms the heuristic baseline for two of the three.

Figure 5.4a shows the results for finding files together. The classifier had an average precision of 69.3%. The mean number of successful similarity recommendations was 2.34 (sd 1.70, min 0, max 7). The mean tree distance of successfully predicted file pairs was 3.47 (sd 2.91, max 21). A promising result is that the classifier outperformed the baseline for deep:far file pairs by 3.9x, reaches a precision of 74%, and predicted successfully at a mean tree distance of 5.63. This can address a common instance in file management: when searching for a file, users navigate to a folder with related content, but which does not contain the file they need. Providing them at that time with the recommendation of the file they are searching for would be highly useful.
Further, the classifier greatly exceeds the heuristic baseline in this instance as well. Figure 5.4b shows the results for moving similar files. The classifier had an average precision of 67.2%. The mean number of successful similarity recommendations it made was 3.28 (sd 2.24, min 0, max 9). The mean tree distance of successfully predicted file pairs was 2.89 (sd 2.29, max 21). We achieve relatively high precision for both flat and deep:close hierarchies (79% and 83%, respectively), and slightly exceed the heuristic baseline in these instances. The result for flat pairs matches expectation: the operation of file movement within limited file hierarchies is not a significant one.

Figure 5.4c shows the results for deleting similar files. The classifier had an average precision of 46.6%. This indicates that classification techniques might not be the correct approach for aiding this file co-management decision. The mean number of successful similarity recommendations by the classifier was 0.93 (sd 0.94, max 4). The mean tree distance of successfully predicted file pairs was 2.13, with a maximum of 7. The most notable result here is that classification on flat pairs beat the baseline by 2.6x and reached a precision of 72%. This also beats the heuristic baseline at 46.7%. This is potentially helpful for users who wish to start using folders. One participant in particular stated that they do not wish to start using folders because of how cluttered their account already is.

Data Similarity’s Correlations With Prediction Tasks

To better understand our classifiers’ high precision compared to the randomized baseline, we further investigated which types of data similarity correlate with participants’ four types of perceived similarity and three file co-management decisions. We created logistic regression models for each of these seven prediction tasks. In each model, the 11 types of data similarity metrics were the independent variables, and the respective perceived similarity or co-management decision (“Strongly agree” and “Agree” binned as 1, all other responses binned as 0) was the dependent variable.
As shown in Table 5.1, many data similarity metrics were predictive for each of the seven classification tasks. No single data similarity metric was sufficient for any task, whereas we found that the combination of metrics to be predictive. For all seven tasks, similarity in a file pair’s last modified dates and filenames were predictive. For six tasks, similarities in file contents and shared users were predictive. In contrast, similarity in text topic and image contents were both predictive of topic similarity and purpose similarity, but not of file co-management decisions.
Chapter 6:
Discussion and Conclusion

Our study validates the use of perceived similarity for organizing and managing file storage, as seen through the lens of cloud storage. In our qualitative evaluation of participants’ responses to our survey, we identified different ways that participants perceived similarity, and found that they generally wished to co-manage files that they perceive as similar. Our classifiers were able to predict this perceived similarity and co-management with high precision compared to a randomized baseline, and match or exceed a heuristic baseline for four of the seven tasks. The precision our classifiers attained is lower than desired for applications of these techniques in the wild, and we believe future work should enhance this signal in order to make these techniques usable.

We envision the techniques we present here as augmenting, rather than replacing, the file hierarchy [47, 36]. Previous techniques [26, 42] that work in this same manner, however, have not been able to incorporate files’ contextual information. This limits their general applicability: when in situations with limited access information, techniques from prior work will likely encounter difficulties. The techniques we propose here do not have such limitations. Further, users are familiar with the hierarchical file system, and techniques that seek to work around it [15, 2, 9, 61] find difficulties.

We view several potential extensions of this work. Firstly, providing more sophisticated techniques on a more targeted set of similarity metrics could enhance the existing recommendations. Here, we simply examined common metric in order to determine feasibility, without examining which would be most effective. We believe techniques such as weak and distant supervision [57, 71] could allow users to provide more sophisticated input, thus improving recommendations. We also view the potential for learning data-specific similarity metrics such as in vector space embeddings or Siamese neural networks to be a potential area of interest [60, 3].

Secondly, examining the scalability of the techniques we present here would provide a valuable research contribution. Investigating whether multimodal similarity hashes [44] can still provide input to our technique could be one extension of this.

Lastly, we view the development of a full tool implementing these techniques as being a critical part of this research path. Our usage of conceptual contextual information could provide enhanced interactions with users beyond what prior work has noted. For example, instead of providing recommendations to users without explanations, a tool based on these techniques could explain, “This other document appears to be an earlier version of the one you just deleted. Do you want to delete this one, and all previous versions?” Or, the techniques could be used to identify
where a newly created file should be saved, therefore preventing disorganization from occurring in the first place. This type of total interaction has not been fully examined in the literature to date.
Chapter 7:

Acknowledgements

Thank you to my advisers, Blase Ur and Aaron Elmore, for all their assistance throughout this process. Thank you also to Kyle Chard, Michelle Aniye, Galen Harrison, and Mohammad Taha Khan, without whom this research would not have been possible.
References


Appendix A: Survey Instrument

Part 1

As part of this study, our computer code collects some data from your Google Drive or Dropbox account. As shown in the image below, it collects some file metadata as well as encrypted text from documents or images and objects recognized in images. It does not save any of the files. Researchers will never have access to the human-readable versions of any file content.

To access these files, we ask you to log in through the secure OAuth service. This provides our code with a one-time access token to use for this study. We do not save any usernames, passwords, or personally identifiable information through this process. If you are interested, you can read more about how Google allows third parties to access your account, and how you can manage this access here: https://support.google.com/accounts/bin/answer.py?hl=en&answer=143031. You may find equivalent information for Dropbox here: https://www.dropbox.com/help/security/third-party-apps.

If these terms are acceptable, please indicate so below. Otherwise, you will be asked to release the submission.

1. I agree to provide access to my Google Drive or Dropbox account under the terms specified above.
   • Yes
   • No

Demographics
1. Are you the only person with access to this account, or do you share the username and password to this account with others?
   - I am the only person with access
   - Other users have access
   - Not sure

2. To the best of your knowledge, how many users (including yourself) have access to this account?

3. For what purpose(s) do you use this account?

4. What percentage of the data in this account would you characterize to be primarily for personal use?

5. What percentage of the data in this account would you characterize to be primarily for professional use? (i.e., related to your job or career)

6. How do you interact with your [Cloud Service] account? Please mark all that apply.
   - I use the website
   - I use the app
   - I sync it with folders on my computer

7. How often do you open the website or app for your [Cloud Service] account?
   - Daily
   - Weekly
   - Monthly
   - Yearly or less
   - I don’t know

8. Please rate your agreement with the following statement: “My [Cloud Service] account is well-organized”
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
   - I don’t know

9. Please describe your strategies for organizing your [Cloud Service] account in 5 or fewer sentences.

10. Do you use folders to organize your account?
• Yes
• No
• I don’t know

11. (If ‘Yes’ to previous question) How?

12. (If ‘No’ to previous question) Why not?

13. Please describe a specific experience in which you were not able to find a file you were looking for in your [Cloud Service] account. If you have not had such an experience, please state so.

14. Please list any other cloud storage services (e.g. Sharepoint, Box, iCloud) that you use personally or professionally.

15. With what gender do you identify?
   • Male
   • Female
   • Non-binary / other
   • I prefer not to answer

16. Are you majoring in, hold a degree in, or have held a job in any of the following fields: computer science; computer engineering; information technology; or a related field?
   • Yes
   • No
   • I don’t know

17. What is your age range?
   • 18-24 years old
   • 25-34 years old
   • 35-44 years old
   • 45-54 years old
   • 55-64 years old
   • 65 years or older
   • Prefer not to answer

18. What is your occupation? (optional)
Part 2

Instructions

Our research goal is to design systems to help people manage their cloud storage accounts. We will present to you 18 pairs of files from your [Cloud Service] account, and ask you to rate their similarity based on the following categories. You will be able to return to these instructions at any point during the survey.

*The Data Itself*

- **Topic**—two files are similar if they talk about the same subject matter
  - Example: a photo of a dog and a document about dog grooming techniques

*Origin*

- **Purpose**—two files are similar if they will likely be used for similar tasks or goals
  - Example: a photo of a dog and a document about dog grooming techniques
- **Derivation**—two files are similar if they are different versions of the same item, or if one “created” the other
  - Example: a rough draft of a proposal document, and a final version of the same document
  - Example: a music score, and a recording of you playing the music from that score

- **Creation Context**—two files are similar if they were created at the same time or in the same place
  - Example: a short story you wrote at a writer’s retreat, and another person’s poem written at the same retreat

*Content*

(The following section is repeated 18 times, each with a different selected file pair)

You are on pair [Current File Pair] out of 18 file pairs. Please answer the following questions in reference to the following two files:

File 1: [File Name A] ([Preview Link]) File 2: [File Name B] ([Preview Link])

Note that you must preview the files at the provided links to continue in the survey. To review the tutorial: [Link to Tutorial]

(Break)

Please answer the following questions in reference to the following two files:

File 1: [File Name A] ([Preview Link]) File 2: [File Name B] ([Preview Link])

To review the tutorial: [Link to Tutorial]

1. Please give a short description of **File 1**
2. Please give a short description of **File 2**
3. Please describe in general how these files are similar or dissimilar.
4. I consider these two files to be similar in Topic.
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
   - I don’t know

5. I consider these two files to be similar in Derivation.
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
   - I don’t know

6. I consider these two files to be similar in Purpose.
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
   - I don’t know

7. I consider these two files to be similar in Creation Context.
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
   - I don’t know

8. It is okay if all copies of File 1 are deleted.
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
9. It is okay if all copies of **File 2** are deleted.
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
   - I don’t know

10. I would be upset if the contents of **File 1** were to be released publicly.
    - Strongly agree
    - Agree
    - Neutral
    - Disagree
    - Strongly disagree
    - I don’t know

11. I would be upset if the contents of **File 2** were to be released publicly.
    - Strongly agree
    - Agree
    - Neutral
    - Disagree
    - Strongly disagree
    - I don’t know

(Break)
Please answer the following questions in reference to the following two files:
File 1: [File Name A] ([Preview Link]) File 2: [File Name B] ([Preview Link])
To review the tutorial: [Link to Tutorial]

12. If I were searching for information, and I found one of these files to be relevant, I would also want to see the other file.
    - Strongly agree
    - Agree
    - Neutral
    - Disagree
    - Strongly disagree
    - I don’t know
13. If I were organizing my [Cloud Service] account, and I wanted to move one of these files to a new location, I would also want to move the other file to that same location.
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
   - I don’t know

14. If I were organizing my [Cloud Service] account, and I wanted to delete one of these files, I would also want to delete the other file.
   - Strongly agree
   - Agree
   - Neutral
   - Disagree
   - Strongly disagree
   - I don’t know

15. Which of the below was/were MOST informative in your answers to the previous 3 questions? (Please mark all that apply.)
   - The similarity of the files’ Topic
   - The similarity of the files’ Derivation
   - The similarity of the files’ Purpose
   - The similarity of the files’ Creation Context
   - I don’t know
## Appendix B: Full Logistics Regression Models

| Coefficient   | Std. Err | z    | P>|z| |
|---------------|----------|------|-----|
| File Contents | 0.993    | 0.339| 2.929| 0.003 |
| Filename      | 1.364    | 0.448| 3.046| 0.002 |
| File Size     | 0.006    | 0.019| 0.297| 0.766 |
| Last Modified | <0.001   | 0.019| <0.001| <0.001 |
| Tree Distance | 0.023    | 0.028| 0.806| 0.420 |
| Shared Users  | 1.136    | 0.330| 3.443| <0.001 |

| Coefficient   | Std. Err | z    | P>|z| |
|---------------|----------|------|-----|
| Text Topic    | 1.138    | 0.376| 3.026| 0.002 |
| Text Contents | <0.001   | 0.335| <0.001| 0.127 |

| Coefficient   | Std. Err | z    | P>|z| |
|---------------|----------|------|-----|
| Tree Distance | 0.023    | 0.037| 0.619| 0.536 |
| Shared Users  | 0.524    | 0.345| 1.517| 0.129 |

| Coefficient   | Std. Err | z    | P>|z| |
|---------------|----------|------|-----|
| Text Topic    | 0.630    | 0.395| 1.597| 0.110 |
| Text Contents | <0.001   | 0.399| <0.001| 0.928 |

| Coefficient   | Std. Err | z    | P>|z| |
|---------------|----------|------|-----|
| Table Schema  | 0.396    | 1.478| 0.268| 0.789 |

| Coefficient   | Std. Err | z    | P>|z| |
|---------------|----------|------|-----|
| Image Color   | 0.002    | 0.004| 0.642| 0.521 |
| Image Contents| 1.879    | 0.749| 2.509| 0.012 |

Table 1: Logistic regression model for **topic** similarity.

Table 2: Logistic regression model for **derivation** similarity.
## Appendix B: Full Logistics Regression Models

|          | Coefficient | Std. Err | z     | P > |t| |
|----------|-------------|----------|-------|-----|---|
| File Contents | 1.499       | 0.341    | 4.399 | <0.001 |
| Filename   | 2.225       | 0.477    | 4.661 | <0.001 |
| File Size  | <0.001      | 0.019    | <0.001 | 0.555 |
| Last Modified | <0.001     | 0.020    | <0.001 | <0.001 |
| Tree Distance | 0.029      | 0.028    | 1.012 | 0.311 |
| Shared Users | 1.220       | 0.348    | 3.502 | <0.001 |
| Text Topic | 1.418       | 0.388    | 3.650 | <0.001 |
| Text Contents | <0.001     | 0.337    | <0.001 | 0.564 |
| Table Schema | 3.471       | 1.405    | 2.470 | 0.014 |
| Image Color | 0.002       | 0.003    | 0.725 | 0.468 |
| Image Contents | 4.510      | 0.948    | 4.756 | <0.001 |

Table 3: Logistic regression model for **pur-pose similarity**.

|          | Coefficient | Std. Err | z     | P > |t| |
|----------|-------------|----------|-------|-----|---|
| File Contents | 0.761       | 0.379    | 2.009 | 0.045 |
| Filename   | 2.548       | 0.452    | 5.639 | <0.001 |
| File Size  | <0.001      | 0.021    | <0.001 | 0.774 |
| Last Modified | <0.001     | 0.020    | <0.001 | <0.001 |
| Tree Distance | 0.045      | 0.030    | 1.510 | 0.131 |
| Shared Users | 0.705       | 0.335    | 2.102 | 0.036 |
| Text Topic | 1.037       | 0.372    | 2.790 | 0.005 |
| Text Contents | <0.001     | 0.363    | <0.001 | 0.088 |
| Table Schema | <0.001      | 1.556    | <0.001 | 0.827 |
| Image Color | 0.002       | 0.003    | 0.544 | 0.586 |
| Image Contents | 1.537      | 0.767    | 2.004 | 0.045 |

Table 4: Logistic regression model for **cre-ation similarity**.
Table 5: Logistic regression model for finding together.

| Coefficient       | Std. Err | z    | P > |t| |
|-------------------|----------|------|-----|---|
| File Contents     | 0.953    | 0.361| 2.641| 0.008 |
| Filename          | 3.234    | 0.473| 6.839| <0.001 |
| File Size         | <0.001   | 0.021| 0.018| 0.986 |
| Last Modified     | <0.001   | 0.020| <0.001| <0.001 |
| Tree Distance     | 0.047    | 0.029| 1.612| 0.107 |
| Shared Users      | 1.354    | 0.327| 4.144| <0.001 |
| Text Topic        | 0.729    | 0.382| 1.908| 0.056 |
| Text Contents     | <0.001   | 0.355| <0.001| 0.203 |
| Table Schema      | <0.001   | 1.593| <0.001| 0.650 |
| Image Color       | <0.001   | 0.003| <0.001| 0.057 |
| Image Contents    | 1.434    | 0.798| 1.798| 0.072 |

Table 6: Logistic regression model for moving together.

| Coefficient       | Std. Err | z    | P > |t| |
|-------------------|----------|------|-----|---|
| File Contents     | 0.773    | 0.345| 2.239| 0.025 |
| Filename          | 3.440    | 0.501| 6.868| <0.001 |
| File Size         | 0.028    | 0.019| 1.444| 0.149 |
| Last Modified     | <0.001   | 0.020| <0.001| <0.001 |
| Tree Distance     | <0.001   | 0.031| <0.001| 0.037 |
| Shared Users      | 1.416    | 0.348| 4.066| <0.001 |
| Text Topic        | 0.287    | 0.379| 0.759| 0.448 |
| Text Contents     | <0.001   | 0.336| <0.001| 0.997 |
| Table Schema      | 0.160    | 1.539| 0.104| 0.917 |
| Image Color       | 0.004    | 0.003| 1.356| 0.175 |
| Image Contents    | 0.825    | 0.771| 1.071| 0.284 |

Table 7: Logistic regression model for deleting together.

| Coefficient       | Std. Err | z    | P > |t| |
|-------------------|----------|------|-----|---|
| File Contents     | 0.411    | 0.398| 1.032| 0.302 |
| Filename          | 1.294    | 0.435| 2.971| 0.003 |
| File Size         | <0.001   | 0.019| <0.001| 0.015 |
| Last Modified     | <0.001   | 0.020| <0.001| <0.001 |
| Tree Distance     | <0.001   | 0.034| <0.001| 0.098 |
| Shared Users      | 0.767    | 0.301| 2.552| 0.011 |
| Text Topic        | 0.174    | 0.381| 0.457| 0.648 |
| Text Contents     | <0.001   | 0.339| <0.001| 0.620 |
| Table Schema      | <0.001   | 1.429| <0.001| 0.630 |
| Image Color       | <0.001   | 0.004| <0.001| 0.076 |
| Image Contents    | 0.288    | 0.701| 0.411| 0.681 |