The Scale and Structure of Personal File Collections

Jesse David Dinneen[*]
Victoria University of Wellington
Wellington, New Zealand
jesse.dinneen@vuw.ac.nz

Charles-Antoine Julien
McGill University
Montreal, Quebec
charles.julien@mcgill.ca

Ilja Frissen
McGill University
Montreal, Quebec
ilja.frissen@mcgill.ca

ABSTRACT

Although many challenges of managing computer files have been identified in past studies – and many alternative prototypes made – the scale and structure of personal file collections remain relatively unknown. We studied 348 such collections, and found they are typically considerably larger in scale (30-190 thousand files) and structure (folder trees twice taller and many times wider) than previously thought, which suggests files and folders are used now more than ever despite advances in Web storage, desktop search, and tagging. Data along many measures within and across collections were log normally distributed, indicating that personal collections resemble imbalanced, group-made collections and confirming the intuition that personal information management behaviour varies greatly. Directions for the generation of test collections and other future research are discussed.

CCS CONCEPTS

• **Human-centered computing** → **Graphical user interfaces**; **Empirical studies in HCI**; • **Applied computing** → *Document management*;

KEYWORDS

personal information management; files; folder navigation

ACM Reference Format:
Jesse David Dinneen, Charles-Antoine Julien, and Ilja Frissen. 2019. The Scale and Structure of Personal File Collections. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300557

1 INTRODUCTION

The task of managing digital files, also called file management (FM), is ubiquitous in computing: at home and at work, people create, download, copy, name, rename, organise (or leave unorganised), delete (or not), share, navigate to, and search for digital files and folders [17]. The result of such activities are, whether created through active collection or passive accumulation [56], collections of files and folders people personally manage for personal or professional use, or *personal file collections*.

Many studies of personal information management (PIM) have furthered knowledge about the user behaviour and difficulties relevant to managing such collections, and many have developed improvements to the interfaces used to manage those collections. So far, however, a confident description of the typical nature of existing collections (e.g., their scale, structure, and contents) has not been established by nor could be inferred from prior studies (e.g., through collation or meta-analysis) [18], and as a result, our knowledge about the phenomenon, and thus our ability to support it, is limited. Because we do not know what users' folder structures are like, nor how files are categorised therein, it is unclear what navigating such structures is like and if the structures resemble (and thus could benefit from studies of) those in other personal contexts [10] or group-made information structures [36]. Further, given the availability of and recent improvements to features like desktop search and file tagging, some have questioned if people really *are* still creating and organising folders [17]. Additionally, because it is unclear what typical file collections are like, it is unclear what representative and commensurable test collections (e.g., used in testing PIM prototype software) should be like [14]. Finally, without a thorough description of the artefact created during the process of managing files, it is difficult to model that process and the many possible determinant factors suggested by past studies, including technological [9, 10], demographic [26, 40], and individual [42, 46] differences.

In other words, a thorough description of the *what* of FM is needed, and would be useful for complementing studies of the *why* by providing a commensurable baseline to aid in interpreting their results; detailed knowledge of both what and why are needed to inform the design of useful improvements to PIM software (e.g., new features and visualisation

[*] Corresponding author

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300557

approaches). Specific questions to answer in alleviating the above issues therefore include:

- Scale: how many files and folders do people keep?
- Structure: how do people arrange folder structures to organise files?
- Structure: how do people organise files into those structures?

This manuscript thus reports an exploratory study of the scale and structure of the collections of 348 participants using Windows, MacOS, and GNU/Linux, and uses select statistical analyses to identify what constitutes a typical collection.

2 RELATED WORKS

Personal information management refers to the study and practice of individuals managing information owned by or about them (i.e., personal information) but also to individuals *personally* managing information (personal or otherwise) [35], typically with the intention of later returning to that information. PIM research has explored a broad range of contexts and activities within this scope, including the common computing phenomenon of managing digital files, where *personal* could mean, for example, using a private computer to backup travel photos, using a company computer to organise project files, or reviewing the scholarly literature previously saved to a laboratory computer.

Hundreds of studies have covered a variety of subtopics within file management [17], including studying people's use of tags [6], searching and navigating to files [7, 53], how people share files [12], and developing augmentations to FM software [23]. To investigate the common PIM categories (or variables) of storing, organising, and retrieving (or exploiting) digital items [5, 35, 58], many studies have characterised the artefacts produced by FM activities: people's digital collections (i.e., files and folders) [17]. Specifically, studies have measured collections' *scale* (i.e., how many files users store) [10, 26], *contents* (i.e., what users store, including file types, file sizes, and file and folder names) [13, 21, 32], *structure* (i.e., how files and folders are organised) [9, 30], and *usage and age* (e.g., how long ago files were created and retrieved) [50, 57].

Personal file collections grow in scale when users actively store files and folders by creating and downloading them, or passively keep (i.e., do not delete) files and folders in downloaded archives or default folders that the operating system provides. The structure of a collection is similarly determined, for example by users' active organising (or meta-level) activities [34, 58] like creating and moving folders and files (e.g., organising files into relevant folders) or passively leaving files and folders wherever they are downloaded. While the scale of collections can be measured simply by counting the items (i.e., files and folders) in a collection [57] or noting its size in bytes [32], measures of structure are more numerous and complicated; definitions of structural terms are provided in Table 1. Structure can be usefully divided into *folder structure*, measured for example as the maximum depth of the folder tree [31] or branching factor [26], and *file structure* (i.e., categorisation), measured, for example, as the average number of files per folder [9] or the average depth of files in the tree [8]. Additional measures and the specific values reported in prior works are provided below when discussing the results of the present study.

Table 1: Definitions of terms describing folder (top) and file (bottom) structure, with the work first using each term.

term	definition
subfolder	A folder within another folder (common term)
root	Any folder that is not a subfolder whether plainly (e.g., C:\ in Windows) or virtually because users cannot or do not access the folders containing it (e.g., /Users/jesse in MacOS) [10]; similar to *locus* [26]
depth	The number of steps required to navigate down to a folder from the root (the root therefore has a depth of 0) [2]
breadth	The number of folders at some depth in a tree [40]
waist	The broadest part (depth) of a tree (new term)
switch	A folder containing one or more folders but no files [2]; also *navigation folder* [9] or *syndetic node* [37]
leaf	A folder containing no folders [2]
branching factor	The mean number of subfolders per non-leaf folder for a tree or depth [26]; alternatively, ratio of files to folders at a depth [2]
file depth	The depth of the folder a file resides in [1]
file waist	The depth with the most files (new term)
unfiled file	A file in a root folder [10]
root pile rate	The percentage of all files that are unfiled [27]
empty folder	A folder containing neither files nor folders [2]

Due to varying study goals, prior works have made differing measures of collections' scale and structure. For example, works discussing collections' scale measure the collection *either* in bytes [32], or total number of folders [10, 28], or total number of files [46], but rarely use all three measures. Similarly, works discussing collections' structure may report the number of files per folder [26], files' depths in the folder tree [22], or the percentage of empty folders [30], but rarely use common measures; none of the measures in these examples are shared across the cited studies. While the variety of reported measures might suggest past results could be collated to form a complete account of personal file collections, this cannot be done meaningfully due to the time span over which studies have taken place (1980 to 2017) and because studied population samples are often incommensurable in nature, size, or both. For example, while some have studied 10-12 academics [25, 40, 57], others have studied 40 engineers [32], or employees at a single software development company [1, 20, though with only one common measure of structure]. Finally, as our analysis below indicates, the statistical techniques used to describe the results of prior works may mean they inaccurately reported some key values about both the scale and structure of observed collections.

We suggest no shortcoming or oversight in prior works, whose objectives differed from those of the present study. Nonetheless, the end result is that there currently exists no coherent, detailed description of the scale and structure of typical personal file collections (i.e., one derived from a relatively large, heterogeneous, non-niche sample, examining files beyond those accessed recently or during a guided tour, and covering all previously used measures), thus leaving unanswered the questions raised in the introduction.

3 METHODOLOGY

There are several methods for collecting data about personal file collections, such as making observations during guided tours [54], recording video of structured tasks [9], and reviewing the outputs of custom-made logging software [57]. Each approach entails advantages and disadvantages [17]; to practically achieve a relatively large sample, we used software designed for this task, called Cardinal [18]. Cardinal is cross-platform and open-source software that integrates and extends the functionality of the various *ad hoc* data collection scripts used in prior studies [10, 28, 40, 46]. In short, Cardinal accesses locations in a participant's folder hierarchy where they have specified they manage files, and while ignoring file contents and not storing names it records a snapshot of various properties of the files and folders (e.g., locations and folder contents) and additional properties not examined here (e.g., file types and sizes). Such functionality is primarily achieved using standard python packages and common system calls (e.g., *os.scandir*, *os.stat*). The justification for Cardinal's creation and its development and prior validation with 46 participants is described in prior work [18]. Before recruiting participants we further tested our implementation of Cardinal (e.g., compiled binaries) by running it in each operating system version and comparing its outputted data to manual counts of each variable (e.g., tree breadth); Cardinal's data always matched our manual counts.[1]

Recruitment and data collection

As our focus is on general computer users managing files, our criteria for participation were only that participants have files stored locally (i.e., not exclusively in the cloud, whether at home, work, or school) that they personally manage, and have the abilities to read English and download and run the software. We recruited participants from February 2016 to August 2018 by posting calls for participation on study recruitment Websites and in online communities on Facebook and Reddit, sending emails to mailing lists (e.g., industrial, governmental, and academic), and contacting colleagues, friends, and family. Participants downloaded the data collection software from our research project Website and ran it on computers where they managed a collection of files. This consisted of answering questionnaires and specifying through a graphical interface where exactly they manage files – we encouraged the inclusion of both active, working areas and backup locations like external drives – and allowing participants to review a summary of the results before choosing to let the software submit the data to the researchers. Participation was therefore remote and anonymous.[2]

Participants' home folders (i.e., potential roots) were suggested by the data collection software as one of several possible locations where they manage files, and participants were encouraged to customise this and add any additional locations (redundancy and false roots were prevented by the software). Any visible and accessible folders and files within such spaces were included in data collection, but locations outside of those specified (i.e., system folders) were not examined, nor were hidden folders (e.g., /Users/jesse/Library in MacOS and /home/jesse/.cache/ in Linux), nor folders that an unprivileged user could not access (e.g., C:\Windows\system32 in Windows, or /bin in MacOS and Linux). In other words, for the present study we operationally define a user's collection as locally stored files (and folders) that the user stores (either explicitly, by downloading, or implicitly, by not deleting) or organises (by arranging, leaving arranged, or not arranging). Examples of such files include those on the Desktop, in My Documents, in Downloads, and any other folders created or placed in the user's home folder.

The collected data comprise text files containing descriptions and representations, in hierarchical JSON format, of the folder structures in portions of users' collections they marked as personally managed (e.g., potentially including external drives, if selected). From these data files we made measures of each collections' scale and structure (e.g., mean number of files per folder within a collection), and then derived descriptive statistics across the sample (e.g., mean number of files per folder across collections' means), to produce 4 measures of scale, 14 measures of folder structure, and 8 measures of file structure (i.e., categorisation). Explanations of how particular measures were calculated are provided where necessary below, and details of how each file system

[1] The authors thank an anonymous reviewer for identifying a bug in Cardinal's interface that inflated the number of empty folders reported to participants. This did not affect the collected data, and has since been corrected.

[2] This study was approved by ethics committees at McGill University (REB #75-0715) and Victoria University of Wellington (HEC #25658).

property is recorded can be found in prior work [18] and in the software's annotated source code[3].

Data analysis

Examination of the collected data (e.g., with plots, tables, and fitting software) revealed all distributions either (a) fit a normal distribution model well or (b) showed positive skew and fit a log-normal distribution model significantly ($p < 0.05$) better than several other similar distributions (e.g., power law, exponential, and negative binomial). For the latter kind of data (i.e., positively skewed and relatively long-tailed), hereafter referred to as log-normal distributions, traditional measures like arithmetic mean do not provide an accurate description of the expected value [43][4], and so for such data we calculated measures designed specifically for log-normal data. Namely, we report a mean defined as $ln(\mu) = e^{\mu+\sigma^2/2}$ and a median and standard deviation derived by log transforming, making traditional mean and SD measures, and back transforming the data (i.e., e^{μ} and e^{σ}, respectively) [43, 49]. The log-normal mean and median thus describe the lower and upper bounds of the range of typical values, respectively, which we refer to as the *typical range*, while the log-normal standard deviation reflects the severity of the skew (e.g., a value of 1.5 indicates a nearly normal distribution with mild right skew, and a value of 7.5 indicates a highly right-skewed distribution that resembles a power law distribution more than a normal one). To differentiate such measures from traditional ones we adopt the labeling used in prior works [43]: mean*, median*, and SD*. To facilitate comparison of our findings with those of prior works we also report the traditional mean and SD of log-normal data, in parentheses.

To reflect typical cases, we removed outliers from the collected data using interquartile range [59], which works for for both normal and log-normal data as it does not require symmetrical distribution [52]. It was also necessary to remove values of zero during derivation of log-normal statistics (because the logarithm of zero is undefined). We note the number of outliers or zero values removed when it is important to accurately reporting or interpreting the results.

4 RESULTS

We received 348 data files, all of which were usable, describing 49.2 million files across 7.9 million folders. Respondents were 60% male and appear heterogeneous, with ages ranging 14 to 64 (mean 30, SD 9.9) and diverse occupations including: poet, marketing director, electronics technician, doctor, bartender, data analyst, product designer, biologist, videographer, safety inspector, and librarian. Data came from a variety of machines (laptops, desktops, and tablets) with varied uses (personal matters, work and/or school, or a combination thereof) and operating systems (Windows XP to 10, MacOS 10.8 to 10.13, and eight Linux distributions). The compositions of these categories and the demographic aspects of the sample are presented in Table 2.

Table 2: Summary of population sample (n = 348) along demographic (top) and technological (bottom) characteristics.

	male	female	other
gender	218 (63%)	123 (35%)	7 (2%)
	range	mean	SD
age	14-64	30	9.9
	MacOS	Windows	GNU/Linux
OS	169 (48%)	135 (39%)	44 (13%)
	laptop	desktop	other (tablet, server)
form	263 (75%)	82 (24%)	3 (1%)
	personal & work/school	work/school	personal
use	254 (73%)	55 (16%)	39 (11%)

Summaries of the collected data are outlined below for each grouping described above (scale, folder structure, and file structure), and tables 3-5 display specific values for each grouping. The implications of our findings, including how they compare to previous findings, are discussed in the next section.

Scale of file collections

The scale of participants' collection varied greatly whether measured by files, folders, or in bytes (SD* values ranging from 6.95 to 7.22, indicating extreme right skew even after outliers were removed). Typical collections, as indicated by the expected values calculated for log-normal measures (i.e., median* to mean*), contain 29 thousand to 193 thousand files (excluding shortcuts, aliases, and symlinks) and 4 thousand to 26 thousand folders. Measured in bytes, collections were typically 33 to 232 GB (mean*). 37 collections (11%) were of outlying scale, with notable instances including (a) a collection of 2.6 million items (2.2 million files and 400 thousand folders; 213 GB) and (b) a collection comprising 4.2 Terabytes (800 thousand files).

Folder structure

Most (60%) collections had one root, indicating that most collections are not split across multiple drives or multiple discrete locations on a single drive. However, roughly 20% of collections had two roots, 9% had three, and 1% had four; these numbers are consistent with the number of hard drives our participants had (as identified by the data collection

[3] http://www.github.com/jddinneen/cardinal

[4] Consider, for example, positively skewed univariate data with values from 0 to 475: the *range* of typical values is better described by the log-normal median of 33 and log-normal mean of 232 than by the traditional arithmetic mean of 92.

Table 3: Measures of the scale of participants' file collections. Collections vary greatly in scale (data in all measures are log normally distributed), but typically contain 29 to 193 thousand files and 4 to 26 thousand folders.

measure	median*	mean*	SD*	(mean; SD)
# files	29,123	193,001	6.99	(73,821; 72,996)
# folders	3,818	26,363	7.14	(10,673; 12,011)
total size (items)	33,900	221,826	6.95	(85,614; 83,756)
total size (GB)	32.92	232.42	7.22	(92.46; 108.99)

software), suggesting that some collections are stored on multiple drives. Immediately under the root level (i.e., the most frequently traversed part) we found a typical range of 15 to 18 folders (SD 1.82), suggesting navigation from the root begins with a decision among that many branches.

Moving down the folder tree, we saw that the waist of most collections was at depth 6.27 (SD 2.41), with the mean depth of all folders nearby at 6.82 (SD 1.79; this excludes twelve outliers, with the most extreme having a mean depth of 12). Typical maximum collection breadth varied greatly, with values ranging from 813 (median*) to 4.4 thousand (mean*) folders broad, and the mean breadth had a typical range from 266 folders (median*) to 853 (mean*). For both the maximum and mean breadth, 10% of participants were extreme outliers (SD* increases from 4.46 to 5.36 if such values are included), with the most extreme having a tree with 276 thousand folders at the waist.

Participants' collections exhibited a mean maximum depth of 15.45 (SD 5.88; i.e., roughly 15 navigation steps are required to navigate from the root to the bottom of the tree; excludes 1 outlying collection with depth of 71), though one participant had a flat collection (i.e., no subfolders). Given a collection of this height, the waist is just past one third of the way down from the root.

At the bottom of collections' tree structures we found a wide typical range (SD* 7.21) of counts of leaves, from 2.6 thousand to 18.2 thousand, which is likely attributable to the widely varying number of folders participants kept, as leaves form the surface area of the bottom of the tree. Leaves accounted for 73% of the folders in participants' collections (SD 7%). 16 participants (4%) had outlying (high) proportions of leaf folders (e.g., 93%) due to their collections being much broader than deep. We observed that the mean depth of leaf folders was near to the waist (6.81), implying not only that the bottom of the tree starts just below the waist but that much of the bottom exist there, such that the tree must taper in breadth towards a relatively narrow maximum depth.

The internal structure of the tree is described by branching factor, which indicates the average number of choices to select from (less one: going upwards towards the root) during a navigation decision within a tree; for example, a branching factor of 3 would mean at any given navigation point (i.e., folder) the tree typically *branches* downwards in three directions, entailing that the typical decision for a user navigating downwards in the tree is between three options. We found collections' branching factors to be 3.62 (SD 0.86), implying the typical downwards navigation decision entails choosing between three or four folders. In this regard there were 35 outlying collections (10%), however, with the most extreme case having a branching factor of 82.6. We found that counts of switches varied greatly, as the counts of all folders did, but that they composed an average of 16% (SD 7%) of participants' folder trees. Five participants (1%) had no switches, and five had outlying, high proportions of switches (e.g., 32%).

Table 4: Measures of the folder structure of participants' collections. Measures of collection breadth are log normally distributed (except branching factor), while roots, proportions, and measures of depth are distributed normally.

measure	median*	mean*	SD*	(mean; SD)
max tree breadth	947	4,990	6.19	(2,482; 2,841)
mean breadth	290	888	4.46	(605; 623)
folders at root	14.88	17.82	1.82	(17.49; 9.99)
leaf folders	2,582	18,192	7.21	(7,183; 7,928)
switch folders	591	4,291	7.33	(1,766; 2148)

	mean (SD)
# roots	1.47 (0.68)
max tree depth	15.45 (5.88)
waist depth	6.27 (2.41)
mean folder depth	6.82 (1.79)
branching factor	3.62 (0.86)
mean leaf depth	7.00 (1.79)
% of leaves	73% (7.0%)
mean switch depth	6.49 (1.75)
% of switches	16% (7.0%)

File structure

We found that 96 collections (28%) had no unfiled files, and those that did typically left only 4 (median*) to 8 (mean*) files unfiled, producing a root pile rate approaching zero. However, 49 collections (14%) had outlying higher numbers of unfiled files, with the most extreme case showing 1,900 unfiled files (nearly 20% of that collection's 10,000 files).

The mean depth of all files within collections was 6.2 (SD 1.7), almost exactly the average tree waist depth, and the file waist was nearby at depth 5.8 (SD 2.4). The typical number of files at the file waist was observed to be 10 thousand (median*) to 52 thousand (mean*), roughly a third of the range of observed collection sizes; depths 5-7 thus contain approximately 33% of a typical collection's files, implying the middle of the tree is the most likely location of a file.

The mean number of files per folder in a collection provides an indication of the difficulty of reviewing and deciding between the items at any given location in a collection, and its distribution provides an indication of the balance or evenness of the classification of items within a collection [36]. Across collections, we found the typical range of number of files per folder (excluding empty folders) was 7 to 8 (SD* 1.7). Within collections, highly uneven file categorisation was observed such that most folders contained few files. While unusually small collections (e.g., 500 files and 100 folders) exhibit only a somewhat right-skewed, log-normal distribution (e.g., SD* of 3.0), larger collections (i.e., the vast majority) exhibit extreme skews resembling a power-law distribution; Figure 1 shows an example of two atypically small collections, the larger of which already demonstrates extreme skew (e.g., SD* of 6.0). In other words, in collections of typical size, most folders contain (or provide access to) relatively few files, while a few folders contain the majority.

Empty folders were also common, typically ranging in count from 300 to 3,000 (SD* 9.18) and comprising 7% to 13% of the entire folder tree (SD* 3.16); eleven participants (3%) had no empty folders, and twenty-nine collections (8%) had outlying, high proportions of empty folders (e.g., 38%).

Table 5: Measures of the file structure of participants' collections. Data along most measures are log normally distributed; two measures of depth are distributed normally.

measure	median*	mean*	SD*	(mean; SD)
unfiled files	4.39	7.55	2.84	(4.96; 6.97)
root pile rate	0%	<0.01%	4.35	(0%; 0%)
files per folder	6.64	7.64	1.70	(7.58; 3.89)
files at file waist	9,892	52,230	6.20	(24,062; 24,324)
empty folders	304	3,057	8.57	(1,018; 1,159)
% empty folders	6%	12%	3.26	(9%; 8%)
	mean (SD)			
mean file depth	6.21 (1.71)			
file waist depth	5.84 (2.37)			

5 DISCUSSION

Scale of file collections

Our participants kept, on average, 29 thousand to 193 thousand files (median* and mean*, respectively). These values are unsurprisingly greater than those reported in the 80's, where participants had 114 files on average [13], though computer scientists had more at 4.5 thousand [51]. The mean number of files people kept grew through the 2000's to between 5 and 8 thousand [26, 30, 32], and more recently to around 15 thousand [46, 57] and even 36.6 thousand [60]. With an arithmetic mean of 73.8 thousand (SD 73 thousand), our data indicate that growth has sustained over time and doubled in size.

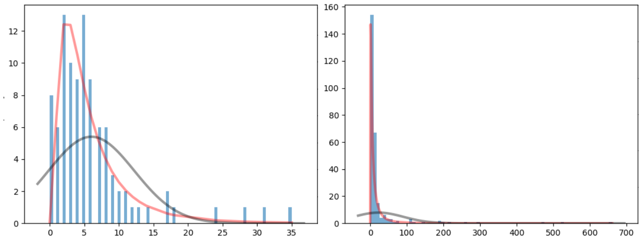

Figure 1: Histograms of two collections, with normal (grey) and log-normal (red) fit lines, showing the highly uneven (i.e., skewed) categorisation of files within folders in the collections observed. The figures show the number of folders (y-axes) containing some number of files (x-axes) in collections with 585 files in 97 folders (left) and 2,035 files in 191 folders (right). The larger collection, though atypically small, already exhibits the highly uneven categorisation seen in the vast majority of participants' collections: most folders contain relatively few files, and a few folders contain most files.

As with files, we observed a variety of folder counts (SD* 7.62) with collections typically having between 4 thousand and 26 thousand folders. This is considerably more than the 6 hundred or fewer found by studies from the 2000's [10, 28, 30, 32] and the 2 thousand seen last year [57]. One study of Microsoft employees [1] saw a mean of 9 thousand folders per collection, approaching the arithmetic mean observed here (11 thousand, SD 12 thousand), but the log-normal mean of 26 thousand suggests it is not unusual for collections to be 3 times bigger still.

To our knowledge is it unknown what portion of files or folders originate locally (i.e., created in a file manager or through 'Save as' dialogues) or are received from external sources (e.g., downloaded in bulk). Regardless of the origins, collections appear to be growing over time, and local storage remains popular despite advances in Web-based file hosting (e.g., Google Drive) and the perception that files are old fashioned [17]. This is perhaps in part attributable to the availability of tools that allow users to store files both remotely *and* locally (e.g., Dropbox), and in part to the relatively low cost of local storage, which provides little incentive to delete files (or move them to exclusively Web-based solutions). Indeed, observed collections typically occupied 33 to 232 GB of storage space; while this is far larger than the largest average collection size reported a decade ago (2.5 GB) [32], the upper bound of 232 GB suggests that most collections can currently be backed up onto a modestly-sized (i.e., 250 GB) external hard drive. These moderate collection sizes also suggest that cases of participants' files being too big to easily back up [41] or transfer between computers [11] are likely atypical. In addition to storage capacity, issues with Web-based storage may discourage users from uploading their collections; despite viewing Web storage like the cloud as a place for personal and shared files [45], users continue to have conceptual issues and interaction difficulties with such storage [45].

Folder structure

The results establish the typical dimensions of the folder trees that make up the organisation of people's file collections, which we describe from the top downwards and illustrate in Figure 2. We observed most collections have a single starting point (i.e., root folder), though some have two, producing a mean consistent with the values of previous studies [26, 32]. From the root folder the tree splits into 15-18 main branches (close to the previously seen average of 19 [40]), and then extends downward, branching at an average rate of 3.62 times the current breadth (SD 0.86). This branching factor is similar to values previously seen in comparable contexts (3.4 and 4.0) [30, 60], and considerably lower than the branching factors previously seen within the first few depths of folder trees (e.g., factors of 8 to 11) [9, 60], suggesting (a) the first navigation decision made when starting from the root thus entails deciding between 10 to 15 more folders (i.e., 3 to 5 times more) than a navigation decision made at any other location and (b) that navigation decisions typically entail fewer options as a user navigates down from the root.

A typical collection then continues expanding at the established branching rate to a broadest point of 1 to 5 thousand folders, entailing it is 300 to 900 folders broad on average. As the tree will typically not extend past a depth of 15 (SD 5.88), it is therefore an order of magnitude broader than deep, contradicting a previous conclusion that trees are narrow (i.e., taller than wide) [26]. This observed maximum depth is almost twice deeper (or taller) than the greatest of the previously reported values, which range from 4.0 to 8.67 [26, 29, 30, 32, 60], perhaps suggesting trees are continuing to grow deeper (or taller) over time. Leaf folders comprise 73% of the collection (SD 7%), only slightly higher than values of 65 to 70% previously observed [1, 20], further indicating that trees have grown broader over time.

With most of the typical collection's external dimensions established, only the vertical location (i.e., depth) of the widest part remains unclear. Folder trees are relatively balanced in this regard, as the mean depth of folders is 6.8 (SD 1.8), near to tree's waist (i.e., depth of 6). This depth is roughly consistent with the previously reported mean folder depths ranging from 5.1 to 6.9 [1, 60]. It is also roughly twice the depth (3.3) reported in one study [10], but as no maximum depth data is reported by [10] it may be that the trees they observed were not particularly deep. It is thus reasonable to conclude that folder trees in personal file collections are roughly symmetrical about their waists, which lay just shy of halfway down the tree. This relative balance is further supported by a previous observation that folder depths within collections are roughly normally distributed [32]. Combined with the observations above (e.g., single root, initial branching of 15-18, maximum width of 900-5000, and maximum depth of 15), we conclude the tree shape is a broad, relatively flat diamond, depicted in Figure 2.

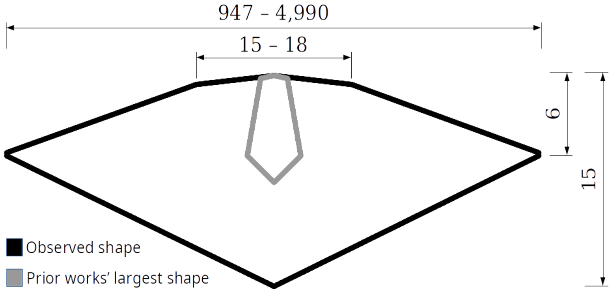

Figure 2: Approximate shape of the folder structure of a typical personal file collection (black) with the largest such shape implied by prior works (grey; 9 folders deep and < 9 broad). Observed collections expand immediately from the root (e.g., Home) folder to 15-18 subfolders, then branch at a rate > 3.62 subfolders per folder down to the waist (~1 to 5 thousand folders broad) at depth 6, where leaves begin to outnumber subfolders as the tree narrows until depth 15.

File structure

We found that almost no files were unfiled (approaching 0% of the collection), which is lower than the values of 2 to 3% seen in prior works [10, 29, 31]. The difference may be attributable to previous studies regarding the desktop [10] and My Documents as pile locations, whereas Cardinal will view such locations as regular folders unless the user specifies they are roots. Whether to regard such folders as piles, and how to define piles (i.e., heaps) [33], remain open questions for PIM research. Regardless, all such values are consistently low, suggesting that while piling is common for paper documents [44] and in digital contexts like emails and Web bookmarks [10] or online learning environments [27], the vast majority of any given file collection is likely to be filed. This may be due to differences in collection types: users may be more invested in organising files, which elicit a sense of ownership [10], and may adapt their organising strategy to the respective, foreseeable retrieval tasks. Definitions of piling that are more inclusive (e.g., regarding piling as placing files in places beyond the root) and further analyses of files *per* depth may reveal more nuanced results, but it is possible that the very presence of so many files necessitates filing simply to maintain the collections' comprehension, accessibility, and navigability [17]. This seems especially likely given the high redundancy of file and folder names (20-30%) observed in prior studies [30, 32], as redundant names increase the difficulty of searching for files by name and thus encourages location-based categorisation into files.

Among files that were categorised into folders (i.e., not piled at the root), we observed collections typically have 7 to 8 files per folder, slightly fewer than the 11 to 13 files per folder reported in studies from 1987 to 2014 [2, 9, 26, 30, 32, 60] and fewer than half as many reported in one study (18.9) [46]. However, if we include outliers and derive an arithmetic mean, as prior studies appear to have done, a similar value results (9.66). Regardless, combining files per folder with the 3.6 folders per folder (i.e., the branching factor) implies 10 to 12 items per folder, consistent with the mean of 11.82 observed in prior work [9], suggesting a stability in categorisation over the last decade. This figure also indicates that users keep fewer items per folder than the number that has been found to cause problems during retrieval by navigation (21) [9]. We also saw that the categorisation of files across folders *within* collections was highly uneven (i.e., most folders contain few files while only a few folders contain most files), as noted above; we discuss this finding at greater length below.

Many empty folders were also seen, with typical collections consisting of 6 to 12% empty folders. This typical range is near to previous values of 8% [30] and 18% [20], suggesting stability of the proportion of empty folders over time and over the collection growth identified here. Such stability agrees with prior work suggesting that collection size has no effect on the number of empty folders [30]. Empty folders may be made, for example, by putting nothing in them at the point of their creation, perhaps in anticipation of forthcoming projects [39, 40], or by not deleting them when the last file or folder is removed. Further work is required to confirm the origin of such folders, clarify their relation to collection size, and understand the impact they have on regular FM tasks.

We observed most files were near depth 6 (±1), though perhaps because of our Windows participants (39%) we also observed slight peaks for files at depths of 2, 3, and 5, which have been identified in some works exclusively studying Windows participants [1, 32]. Surprisingly, however, we did not find the reported Windows peaks at depths of 4 [9] and 7. Studies observing only recently accessed files found lower mean depths (ranging from 1.81 to 3.7) [8, 9, 22], suggesting that files at or deeper than the tree waist or file waist are likely accessed less frequently. Storing less-frequently accessed files deeper in the tree may be an adaptive behaviour to aid navigating back to them by using the descriptive reminders provided by folder names about the content they contain [3]. Regardless, the stark difference between apparent *archive* and *active-use* areas of collections suggests a need to consider both parts of collections, together and separately, in future works, for example, by analysing file access times and patterns across folder tree depths.

Synthesis of folder and file structures. We have identified that typical users keep thousands of folders in their personal file collections, allocating 14% of their collections to organisation, 16% of which are used exclusively for categorising additional folders or aiding navigation (i.e., switches) and 6-12% of which are empty (i.e., for future organisation), and that almost all files organised into folders rather than piled at the root. While it is unclear what proportion of collections' folders were acquired in bulk downloads, they nonetheless can be considered as being used to manage a collection: if downloaded, in their new home they will either be re-organised by the new user or left in the arrangement provided by the original user. To our knowledge, no prior studies suggest what proportion of users' collections come from bulk downloads. Further, most of our categorisation results are consistent with those of prior works, suggesting stability in how files have been categorised over time. We therefore conclude that despite advances in features like desktop search and tagging, people are still typically using folders to organise their collections, and organising them roughly the same as when collections were half the size and smaller.

Distributions within and across collections

Fifteen of our 26 measures (62%; all 4 measures of scale and 11 of 22 measures of structure) were positively skewed, better fitting log-normal distributions than normal and other positively skewed distributions. A high-level interpretation of this could be that it reflects the frequent observation in PIM studies that PIM activity is highly personal and therefore varied [35, 42]; this appears as true of file management as any other PIM activity.

What could be the cause of this distribution? Log-normal distributions are common in empirical data and reflect that the underlying processes are multiplicative in nature [43], and there is some prior evidence of multiplicative file management actions: for example, the multiplicative nature of copying and pasting may be causing file size distributions to appear log normal [21] and file type (i.e., extensions) distributions to follow a power law [15], which is even more extreme. Though these examples are from studies of disk usage, something similar could be happening with scale and structure; copying and pasting folders grows a collection, and exaggerates the already broader-than-tall folder structure and any existing unevenness in the categorisation of files therein. Such actions may later cause further unevenness; for example, as the folder tree grows it may become more difficult to find any particular folder, and so the most findable ones may receive the majority of new files. Future works may verify these ideas by identifying any multiplicative actions frequently performed by users and the amount of time they have had to allow such actions to have exponential effects in a collection.

There are also methodological implications of the observed distributions. We note that previous studies of collections' scale and structure have seen indications of such distributions, for example in high standard deviations [9, 30, 32], the need to normalise (e.g., log transform) data to prepare it for statistical tests that assume normality [46], and the usefulness of visualising some data with cumulative distribution plots [1]. Nonetheless, such studies have described their data with traditional statistics (e.g., arithmetic mean) and in doing so have likely underestimated the range of typical values in their data, and thus possibly misrepresented collections' scale and structure (or the behaviour that produced them). It is therefore advisable for future FM studies – and any PIM studies examining similar digital items – (1) to check for positive-skew distributions (e.g., log-normal, power law, etc), (2) consider if it would be helpful to describe the data with distribution-specific statistics (e.g., median* and mean*), and (3) apply any appropriate transformations (e.g., negative binomial modeling) [48] before performing statistical tests that assume normality.

Similarity to group-made information structures

Identifying similarities between information structures is desirable because it may provide insights into the creation and use of the structures, as well as direction for possible research approaches to adapt from one context to another. We are not aware of any information structures that resemble the shape of personal file collections as evidenced by the folder tree (e.g., Figure 2). However, as we note above, files in such collections are distributed (or categorised) across folders in a highly uneven manner. Specifically, as collection size increases the unevenness of the files-per-folder distribution appears to increase, evolving from log-normality to appearing to approach a power law distribution (shown in Figure 1). In this regard, personal file collections resemble several large information structures [4] made not by individuals but teams of experts, including controlled vocabularies like the Library of Congress Subject Headings (LCSH) and Medical Subject Headings (MeSH) [36].

The similarity of personally-made structures to group-made structures is notable as it suggests potentially helpful analyses and interventions applied to group-made structures could be adapted to folder structures. For example, a measure (called browsing complexity) has been derived for quantifying the difficulty of navigating group-made subject heading trees [36]; the similarities seen here suggest such a measure could reasonably be adapted to analyse and compare the difficulty of navigating personal collections (e.g., folder trees). In other words, PIM researchers may have a ready-made measure of the difficulty of navigating particular folder structures, which can be used, for example, to understand the difficulty that a folder structure contributes to particular file management tasks (e.g., refinding). Additionally, recent studies have demonstrated the promise of automated modifications to group-made structures, for example exploiting uneven categorisation to simplify [36, 38] trees and facilitate users performing retrieval tasks with them [16]. As the folder trees in personal file collections appear to be similar (e.g., uneven in the distribution of their contents and containing a high proportion of empty folders), such automated approaches could be usefully adapted to personal collections. This could be done, for example, through implementing real-time hiding of folders by adding to the file manager a slider bar that controls how full a folder must be to remain visible [38], effectively simplifying the display of a folder tree, which may facilitate a new employee in comprehending their predecessor's work files or retrieving files from an organisation's shared drive.

Implications for generating test collections

We suggest the present findings should be reflected in generated test collections (or selected existing collections) used for testing PIM software (e.g., prototypes and augmentations) [14], including, for example, the range of typical values (i.e., log-normal distributions) and the uneven distribution of files across folders. This is especially important when it matters that a collection have the scale and structure representative of typical collections.

The range of values seen here can be reflected in test collections, whether generated or selected, in various ways. For example, when using only a single collection, its scale and structure could be within the typical values identified here (e.g., median* and mean*, or near to the arithmetic mean, as appropriate per measure), or alternatively, multiple collections could be created to represent the lower and upper ends of the typical range of values (e.g., one using the median* values and one using the mean* values). The specific values provided here should therefore constitute a good starting point for generating or selecting collections that have the scale and structure typical of existing personal file collections.

6 LIMITATIONS

Some notable limitations to our findings are inherent to our data collection method. Data were gathered in single snapshots of users' collections, and only on their local machines, and thus give no indication of traceless actions (e.g., deleting, renaming, sharing, navigating, or searching files) and do not describe collections stored exclusively in the cloud. Data were also collected without user annotation (e.g., reflections or rationale), and thus provide limited evidence of what users experienced while creating or maintaining the collections. Future studies may take observational and/or longitudinal approaches (e.g., using repeated snapshots or logging software) [19] to see specific actions and changes in the collections over time, and might use cloud platforms' APIs to collect data about files in such locations. Finally, the data collection software used also defines piles strictly, as discussed above, and so studies using more flexible definitions may find piling is closer to 3%, as in prior work.

Additional limitations result from the data reported and analyses performed. We examined here only the scale and structure of people's file collections, rather than the *contents*, which are necessary to form a complete picture of the collections and the activities that produced them and, depending upon studies' exact goals, may be necessary when generating or selecting representative test collections. We therefore encourage future works to examine, for example, files' types, names, and ages. The results presented here are also purely descriptive; we have not attempted, for example, to determine the effects of (or differences across) operating system, collection age, or users' individual differences, as prior works have suggested [9, 46, 55]. We hope the measures, analyses, and findings presented here are can be helpful in investigating such effects in future works and contextualising their results.

Although data were found to be closest to normal or log-normal distributions (e.g., preferable to several similar distributions), the list of distributions tested was not exhaustive and so some distributions may fit the skewed data better (e.g., double-pareto, Weibull, Rayleigh, Poisson) [47] and may provide more descriptive measures of typical values. Nonetheless, the log-normal measures provided are an improvement over traditional measures (e.g., arithmetic mean and SD) for describing the distributions observed. Finally, the analyses performed in this study were intended to describe typical cases, and thus we removed outlying values. While outliers were infrequent, giving them consideration may provide insights into extreme user behaviour, and so they deserve further attention in future studies.

7 CONCLUSION

Understanding and supporting file management requires understanding its activity and the artefact it produces: personal file collections. We have provided here a description of the current scale and structure of such collections, and thus a common point of comparison across prior, disparate works. We identified changes in collection scale and structure over time, suggested how to provide typical values despite the variety seen in collections, and indicated how test collections can reflect that variety.

Our findings suggest that computer users still use (or at least keep) many files and folders, and although typical collections are growing in scale (up to nearly 200,000 files) and external structure (folder trees sized 15 x 5,000), the categorisation of files within remains stable. That folders are still kept supports an argument advanced by [17]: as a collection grows, categorisation becomes necessary to keep it comprehensible, navigable, and accessible, and folders support this need by categorising the many files people are storing. That categorisation within collections is highly uneven demonstrates personal collections resemble group-made information structures, and suggests it may be possible to adapt improvements from the interfaces for such structures to FM software.

While the goal of this study was to establish what collections people have created, rather than why they did it or how to improve the relevant interactions, our findings can contribute to studies posing such questions. For example, the tree depths observed suggest it would be useful to further develop navigation aids [24, 55], while studies of particular populations' FM [32, 40] and other PIM collections (e.g., email and Web bookmarks) now have a general context to which those results can be compared. Future work will present the results of collections' contents and analyse differences across factors like operating system, user occupation, and individual differences.

While storing hundreds of thousands of files and classifying them into several thousand folders may sound like an extreme case of information management, our results suggest typical computer users are in fact doing this. We believe quantifying this phenomenon was a necessary step towards further study of FM, such as user modelling and developing standardised test collections. File management software was first designed when people kept very few files and folders on shared workstations and managed them with textual commands, and its modern, graphical counterparts have offered a stable core of functionalities since their early versions from the 1980's despite collection sizes growing in orders of magnitude. As our knowledge of FM improves, so too can FM software and services improve in their ability to support users managing and browsing large, personal collections.

ACKNOWLEDGMENTS

The authors thank the anonymous CHI reviewers for their helpful feedback on drafts of this manuscript, and the many anonymous participants for their time.

REFERENCES

[1] Nitin Agrawal, William J Bolosky, John R Douceur, and Jacob R Lorch. 2007. A five-year study of file-system metadata. *ACM Transactions on Storage (TOS)* 3, 3 (2007), 9.

[2] Omer Akin, Can Baykan, and D Radha Rao. 1987. Structure of a directory space: A case study with a UNIX operating system. *International Journal of Man-Machine Studies* 26, 3 (1987), 361–382.

[3] Deborah Barreau and Bonnie A Nardi. 1995. Finding and reminding: file organization from the desktop. *ACM SIGCHI Bulletin* 27, 3 (1995), 39–43.

[4] Marcia Jeanne Bates. 2003. *Task force recommendation 2.3 research and design review: improving user access to library catalog and portal information: final report (version 3).* Library of Congress.

[5] Ofer Bergman. 2013. Variables for personal information management research. In *Aslib Proceedings*, Vol. 65. Emerald Group Publishing Limited.

[6] Ofer Bergman, Noa Gradovitch, Judit Bar-Ilan, and Ruth Beyth-Marom. 2013. Folder versus tag preference in personal information management. *Journal of the American Society for Information Science and Technology* 64, 10 (2013), 1995–2012.

[7] Ofer Bergman, Maskit Tene-Rubinstein, and Jonathan Shalom. 2013. The use of attention resources in navigation versus search. *Personal and Ubiquitous Computing* 17, 3 (2013), 583–590.

[8] Ofer Bergman, Steve Whittaker, and Noa Falk. 2014. Shared files: The retrieval perspective. *Journal of the Association for Information Science and Technology* 65, 10 (2014), 1949–1963. https://doi.org/10.1002/asi.23147

[9] Ofer Bergman, Steve Whittaker, Mark Sanderson, Rafi Nachmias, and Anand Ramamoorthy. 2010. The effect of folder structure on personal file navigation. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2426–2441.

[10] Richard Boardman and M Angela Sasse. 2004. Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *CHI '04 Proceedings of the SIGCHI Conference on Human factors in Computing Systems.* ACM, 583–590.

[11] Robert Capra. 2009. A survey of personal information management practices. In *Personal Information Management workshop at ASIS&T '09: Annual Meeting of the Association for Information Science and Technology.* Vancouver, British Columbia, Canada.

[12] Robert Capra, Emily Vardell, and Kathy Brennan. 2014. File synchronization and sharing: User practices and challenges. *ASIS&T '14: Proceedings of the 77th Annual Meeting of the American Society for Information Science and Technology* 51, 1 (2014), 1–10. https://doi.org/10.1002/meet.2014.14505101059

[13] John M Carroll. 1982. Creative names for personal files in an interactive computing environment. *International Journal of Man-Machine Studies* 16, 4 (1982), 405–438.

[14] Sergey Chernov, Gianluca Demartini, Eelco Herder, Michał Kopycki, and Wolfgang Nejdl. 2008. Evaluating personal information management using an activity logs enriched desktop dataset. In *Personal Information Management workshop at CHI '08: ACM CHI Conference on Human Factors in Computing Systems.* Citeseer.

[15] Yuya Dan and Takehiro Moriya. 2011. Analysis and Simulation of Power Law Distribution of File Types in File Sharing Systems. In *SIMUL '11: The Proceedings of the Third International Conference on Advances in System Simulation.* IARIA.

[16] Jesse David Dinneen, Banafsheh Asadi, Ilja Frissen, Fei Shu, and Charles-Antoine Julien. 2018. Improving Exploration of Topic Hierarchies: Comparative Testing of Simplified Library of Congress Subject Heading Structures. In *CHIIR '18: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval.* ACM, 102–109.

[17] Jesse David Dinneen and Charles-Antoine Julien. 2019. The ubiquitous digital file: a revew of file management research. Author preprint. Retrieved January 7, 2019 from http://staff.sim.vuw.ac.nz/jesse-dinneen/papers/FM-review-preprint.pdf

[18] Jesse David Dinneen, Fabian Odoni, Ilja Frissen, and Charles-Antoine Julien. 2016. Cardinal: novel software for studying file management behaviour. In *ASIST '16: Proceedings of the 79th Association for Infomation Science & Technology Annual Meeting.* Copenhagen, Denmark, Article 62, 10 pages.

[19] Jesse David Dinneen, Fabian Odoni, and Charles-Antoine Julien. 2016. Towards a desirable data collection tool for studying long-term PIM. In *Personal Information Management workshop at CHI '16: ACM CHI Conference on Human Factors in Computing Systems.* ACM. http://pimworkshop.org/2016/papers/PIM_2016_paper_16.pdf

[20] John R Douceur and William J Bolosky. 1999. A large-scale study of file-system contents. *ACM Performance Evaluation Review (SIGMETRICS)* 27, 1 (1999), 59–70.

[21] Allen B Downey. 2001. The structural cause of file size distributions. In *MASCOTS '01, Proceedings of the Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems.* IEEE, 361–370.

[22] Stephen Fitchett and Andy Cockburn. 2015. An empirical characterisation of file retrieval. *International Journal of Human-Computer Studies* 74 (2015), 1 – 13.

[23] Stephen Fitchett, Andy Cockburn, and Carl Gutwin. 2013. Improving navigation-based file retrieval. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2329–2338.

[24] Stephen Fitchett, Andy Cockburn, and Carl Gutwin. 2014. Finder highlights: field evaluation and design of an augmented file browser. In *CHI '14: Proceedings of the 32nd Annual ACM CHI Conference on Human Factors in Computing Systems.* ACM, 3685–3694.

[25] Daniel J Gonçalves and Joaquim A Jorge. 2003. Analyzing personal document spaces. In *HCII '03: Proceedings of the 10th annual conference of Human-Computer Interaction International*, Vol. 24.

[26] Daniel J Gonçalves and Joaquim A Jorge. 2003. An empirical study of personal document spaces. In *Interactive Systems. Design, Specification, and Verification. DSV-IS 2003. Lecture Notes in Computer Science.* Vol. 2844. Springer, 46–60.

[27] Sharon Hardof-Jaffe, Arnon Hershkovitz, Hama Abu-Kishk, Ofer Bergman, and Rafi Nachmias. 2009. How Do Students Organize Personal Information Spaces? *International Working Group on Educational Data Mining* (2009).

[28] Sarah Henderson. 2005. Genre, task, topic and time: facets of personal digital document management. In *NZCHI '05: Proceedings of the 6th ACM SIGCHI New Zealand Chapter's International Conference on Computer-Human Interaction.* ACM, 75–82.

[29] Sarah Henderson. 2011. Document duplication: How users (struggle to) manage file copies and versions. In *ASIS&T '11: Proceedings of the American Society for Information Science and Technology Annual Meeting*, Vol. 48. Wiley Online Library, 1–10.

[30] Sarah Henderson and Ananth Srinivasan. 2009. An empirical analysis of personal digital document structures. In *Human Interface and the Management of Information. Designing Information Environments. Lecture Notes in Computer Science, vol 5617.*, M J Smith and G Salvendy (Eds.). Springer, Berlin, Heidelberg, 394–403.

[31] Sarah Henderson and Ananth Srinivasan. 2011. Filing, piling & structuring: strategies for personal document management. In *HICSS '11: 44th Hawaii International Conference on System Sciences.* IEEE, 1–10.

[32] Ben J Hicks, Andy Dong, R Palmer, and Hamish C McAlpine. 2008. Organizing and managing personal electronic files: A mechanical engineer's perspective. *ACM Transactions on Information Systems (TOIS)* 26, 4 (2008), 23.

[33] Dominic Hyde and Diana Raffman. 2018. Sorites Paradox. In *The Stanford Encyclopedia of Philosophy* (summer 2018 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[34] William Jones. 2007. Personal information management. *Annual Review of Information Science and Technology* 41, 1 (2007), 453–504.

[35] William Jones, Jesse David Dinneen, Robert Capra, Manuel Pérez-Quiñones, and Anne R Diekema. 2017. Personal Information Management (PIM). In *Encyclopedia of Library and Information Sciences* (4th ed.), J.D. McDonald and M. Levine-Clark (Eds.). CRC Press.

[36] Charles-Antoine Julien, Pierre Tirilly, Jesse David Dinneen, and Catherine Guastavino. 2013. Reducing Subject Tree Browsing Complexity. *Journal of the American Society for Information Science and Technology* 64, 11 (2013), 2201–2223.

[37] Charles-Antoine Julien, Pierre Tirilly, John E Leide, and Catherine Guastavino. 2012. Constructing a true LCSH tree of a science and engineering collection. *Journal of the American Society for Information Science and Technology* 63, 12 (2012), 2405–2418.

[38] Charles-Antoine Julien, Pierre Tirilly, John E Leide, and Catherine Guastavino. 2012. Exploiting major trends in subject hierarchies for large-scale collection visualization. In *Visualization and Data Analysis 2012*, Vol. 8294. International Society for Optics and Photonics, 82940Z.

[39] Azrina Kamaruddin, Alan Dix, and David Martin. 2006. Why do you make a folder?. In *Adjunct proceedings of HCI2006: The 20th BCS HCI Group Conference*. British Computer Society.

[40] CS Khoo, Brendan Luyt, Caroline Ee, Jamila Osman, Hui-Hui Lim, and Sally Yong. 2007. How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour. *Information Research* 12, 2 (2007), paper 293.

[41] Matjaž Kljun, John Mariani, and Alan Dix. 2016. Toward understanding short-term personal information preservation: a study of backup strategies of end users. *Journal of the Association for Information Science and Technology* 67, 12 (2016), 2947–2963.

[42] Mark W Lansdale. 1988. The psychology of personal information management. *Applied Ergonomics* 19, 1 (1988), 55–66.

[43] Eckhard Limpert, Werner A Stahel, and Markus Abbt. 2001. Log-normal distributions across the sciences: Keys and clues. *BioScience* 51, 5 (2001), 341–352.

[44] Thomas W Malone. 1983. How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems (TOIS)* 1, 1 (1983), 99–112.

[45] Cathy Marshall and John C Tang. 2012. That syncing feeling: early user experiences with the cloud. In *DIS '12: Proceedings of the Designing Interactive Systems Conference*. ACM, 544–553.

[46] Charlotte Massey, Sean TenBrook, Chaconne Tatum, and Steve Whittaker. 2014. PIM and personality: what do our personal file systems say about us?. In *CHI '14: Proceedings of the 32nd Annual ACM CHI Conference on Human Factors in Computing Systems*. ACM, 3695–3704.

[47] Michael Mitzenmacher. 2004. Dynamic models for file sizes and double pareto distributions. *Internet Mathematics* 1, 3 (2004), 305–333.

[48] Robert B O'hara and D Johan Kotze. 2010. Do not log-transform count data. *Methods in Ecology and Evolution* 1, 2 (2010), 118–122.

[49] TB Parkin and JA Robinson. 1992. Analysis of lognormal data. In *Advances in Soil Science*. Springer, 193–235.

[50] Pamela Ravasio, Sissel Guttormsen Schär, and Helmut Krueger. 2004. In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 11, 2 (2004), 156–180.

[51] Mahadev Satyanarayanan. 1981. A study of file sizes and functional lifetimes. *ACM Operating Systems Review (SIGOPS)* 15, 5 (1981), 96–108.

[52] Songwon Seo. 2006. *A review and comparison of methods for detecting outliers in univariate data sets*. Ph.D. Dissertation. University of Pittsburgh.

[53] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 415–422.

[54] Leslie Thomson. 2015. The guided tour technique in information science: explained and illustrated. In *ASIS&T '15: Proceedings of the 78th ASIS&T Annual Meeting of the Associations for Information Science and Technology*. St. Louis, Missouri, 135:1–135:5.

[55] Kim J Vicente, Brian C Hayes, and Robert C Williges. 1987. Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors* 29, 3 (1987), 349–359.

[56] Rebecca D Watkins, Abigail Sellen, and Siân E Lindley. 2015. Digital collections and digital collecting practices. In *CHI '15: Proceedings of the 33rd Annual ACM CHI Conference on Human Factors in Computing Systems*. ACM, 3423–3432.

[57] Roger Whitham and Leon Cruickshank. 2017. The Function and Future of the Folder. *Interacting with Computers* 29, 5 (2017), 1–19.

[58] Steve Whittaker. 2011. Personal information management: From information consumption to curation. *Annual Review of Information Science and Technology* 45 (2011), 1–62.

[59] Rand Wilcox. 2011. *Modern statistics for the social and behavioral sciences: A practical introduction*. CRC press.

[60] Hong Zhang and Xiao Hu. 2014. A quantitative comparison on file folder structures of two groups of information workers. In *JCDL '14: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. 485–486.