Social Influences on User Behavior in Group Information Repositories

by

Emilee Jeanne Rader

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in The University of Michigan
2009

Doctoral Committee:

Professor Judith Spencer Olson, Co-Chair
Research Associate Professor Stephanie Teasley, Co-Chair
Professor Fiona Lee
Associate Professor Soo Young Rieh

"Because there is a group of us using the site I find it difficult to keep up with what is located where, and so ask people to email me direct copies of materials I need."

— a group information repository user, the inspiration for this research

For Dr. Bethany, because she finished first.

Acknowledgments

It is truly mind-boggling to think about the enormity of the investment of time and resources necessary for earning a doctoral degree. I could not have reached this point without the generous financial support of the School of Information and the Rackham Graduate school.

I thank Judy Olson for taking a chance on me by giving me the opportunity to study here. She helped me transition from an industry research lab to the culture of academia, and guided me towards becoming the scientist I always wanted to be. I also thank Stephanie Teasley for stepping as my advisor during my final year, and for finding a way to keep paying me so I could finish.

The BlearyTheory lab group provided emotional support and a place to vent, a venue for engaging intellectual discussion, and great patience and feedback for all of my not-quite-ready-for-prime-time practice talks. I couldn't have done it without you guys. Also, the assistance of the anonymous people who participated in this research was essential to its completion. We call them "subjects" because of the weird questions and tasks we ask of them (or subject them to), and I am sure the research in this thesis seemed strange enough at times. But without them, there would have been no data to keep me up late at night.

I thank my parents for supporting me through difficult times, and for encouraging me believe that I could be anything I wanted to be when I grew up. It is not a coincidence that my sister and I are both researchers in fields where men typically outnumber women. I would not be here today if my mom was the kind of person who was squeamish about picking up a cool bug to get a closer look, or if my dad hadn't let me stay up past my bedtime to watch Doctor Who when he wasn't supposed to.

Finally, I thank my husband Rick Wash for generously building the system used in my experiments, for helping me become a better programmer, and for countless late-night conversations about statistics and other nerdy topics (like whether psychologists and economists are really studying the same things or not). This thesis would have ended up looking very different without his love and support.

Table of Contents

List of Tables

Table

List of Figures

Figure

List of Appendices

Appendix

Abstract

Group information repositories are systems for organizing and sharing files kept in a central location that all group members can access. These systems are often assumed to be tools for storage and control of files and their metadata, not tools for communication. The *storage* approach focuses on providing users with detailed information about the objects in the system—where they are, which users have been looking at them, how they've been used in the past, etc. However, group information repositories tend to grow and become disorganized over time, such that users have difficulty finding what they need. A different approach is to think of these systems as *social* tools that could be governed by the same processes as face-to-face communication, like grounding and audience design.

The purpose of this research is to better understand user behavior in group information repositories, and to determine whether social factors might shape users' choices when labeling and organizing information. While the functionality and capabilities of these systems are essentially the same as the desktop metaphor of personal information management (PIM) systems, I argue that social pressures and processes affect the information structure of the repository, and how it grows and evolves over time. Through a series of interviews with users of a typical group information repository system and an analysis of system log data, I found that users tend to restrict their activities in a repository to files they "own," are reluctant to delete files that could potentially be useful to others, dislike the clutter that results, and can become demotivated if no one views files they uploaded.

I also conducted a two-part online experiment in which participants labeled and organized short text files into a file-and-folder hierarchy. Eighty-four participants were recruited from two intellectual communities (41 Computer Science graduate students, and 43 Information Science graduate students), such that some participants would share community membership common ground with each other, and some would not. Participants were instructed to organize the files for one of three different audiences: themselves, someone from the same intellectual community, and someone from the other community. Forty-eight participants returned four to six weeks later and completed a series of search tasks, in which they browsed hierarchies created by other participants to find specific files. Including

both labeling/organizing and finding tasks in the experiment allowed me to detect potential performance differences when participants searched hierarchies created by others from the same community (or not), and tailored for different audiences. I found that when participants created hierarchies for an audience they imagined was like them, everyone found files in fewer clicks, regardless of whether they were from the same community as the person who created the hierarchy. Further, quantitative analyses of three aspects of the hierarchies (topology, vocabulary, and semantics) helped to explain these results. Users performed better when file and folder labels were more similar to the text of the documents they represented; this correlation was significantly stronger when participants organized the documents for someone who was similar to them.

These results confirm that *audience design*, a communication process, can in fact impact group information management tasks. The findings from both studies suggest that sharing files via a group information repository is more complicated than simply making them available on a server so that others might access them. My research indicates that processes which have been shown to affect spoken communication also impact word choices when the "interaction" is mediated by a repository. Social factors affect users' choices regarding how files in the repository are organized and labeled and what information is retained over time; this in turn affects access to information. Knowing that repositories are social systems will allow system designers to incorporate information that makes the users more salient and familiar to each other, so the process of negotiating shared meaning is better supported by the repository system.

Chapter 1

Introduction

Consider the following examples of online information sharing and reuse:

- A scientist needs to locate some procedures and results from an experiment conducted by another researcher in his lab.
- A student learning the open-source, command-line statistical computing environment R needs to find out how to calculate the mode of her dataset.
- A new member of a design team needs to review requirements analysis activities that took place before he joined the team.
- An intelligence analyst needs to consult information collected by other agencies to assess a potential threat.

Finding the information one needs in situations like these is not straightforward. The scientist looking through someone else's experiment procedures and data encounters information that may not be fully documented or organized in a way that makes sense to her. The student learning R becomes frustrated when her search for the statistical mode function fails; while a function called "mode" exists, it doesn't actually calculate the mode[1]. The new design team member must navigate a vast intranet repository of documents and artifacts generated through the requirements-gathering process, lacking the context necessary to identify what might be useful. And the intelligence analyst must solve an information puzzle, with pieces scattered across agencies having differing priorities and protocols, and using different vocabulary for the same kinds of things.

Despite differences in specific details, these situations have four things in common. First, an *information consumer* must locate information that someone else—an *information producer*—created and shared online. Second, *sharing* means posting information to a shared blog, contributing to a wiki, or uploading a file to a shared folder; the information is made

[1] http://tolstoy.newcastle.edu.au/R/e6/help/09/01/2475.html

available online without specifying a particular recipient (Rader, 2009; Voida, Edwards, Newman, Grinter, & Ducheneaut, 2006). Third, the information that is shared is *explicit*: it has already been captured or documented in some external, concrete way. And fourth, when information producers contribute to a group information system they must *package*, or label and organize the information for others to use; said another way, packaging is the work information producers do that enables a future information consumer to locate and make sense of the information. Effective packaging is not easy; it requires that producers be aware of the knowledge, information needs, expectations and context of future consumers who might need the information (Markus, 2001).

In face-to-face communication, speakers automatically tailor their utterances for their listeners in an effort to ensure that they are able to converge on a shared meaning and understand one another; this is referred to as *audience design*. In an information sharing system, information producers, who contribute information, must similarly *package*, or encapsulate and structure that information for consumers who access and use it. Packaging is the work producers do that enables consumers to find, understand and use the information. Effective packaging is inherently social; it requires that information producers consider both their own ideas and assumptions related to the information objects, and what they know about the information needs and context of whomever might want to find and access the information. What producers and consumers know about each other, if anything, usually comes from sources outside the system; but, if users were given the right information and feedback at the right time, they might be able to communicate better *through* the system.

Information sharing systems are often viewed as storage media, rather than social media; in this dissertation I argue that these systems must be re-conceptualized as a form of asynchronous communication between information producer and consumer. The system mediates this communication, linking producers and consumers via the process of packaging information for reuse, and by the way the information objects are labeled and organized in the system.

By shifting this focus I am able to draw upon what is already known about how people reach common understanding and shared meaning in other kinds of circumstances. For example, the "social" perspective highlights language use as important, because words are chosen to serve as a *handle or identifier* for the information. Language comprises the infrastructure by which information in a given system is found and accessed. Deciding what to name a file and what folder to put it in are packaging decisions, and they are also choices that constrain others' future use of that file as the contents of an information sharing system grow and evolve over time (Rader, 2009).

1.1 Group Information Repositories

Many different types of information sharing systems and services are available to end users. Some examples include: content management systems like Drupal or OpenText Livelink, blogging platforms like Wordpress, version control systems like Subversion, course management systems like Sakai, document management systems like GoogleDocs, collaboration support systems like Microsoft Sharepoint, operating-system-based shared network folders, etc. A full classification and description of all the systems people might use to share information online is outside the scope of this document. However, the research presented here deals with functionality that many of these information systems have in common: the group information repository. When I refer to "group information repositories" in this document I specifically mean the ability to share files via a central, online location with a specific group of people. This centralized storage space is often represented and manipulated using the familiar file-and-folder desktop metaphor that has been used by knowledge workers for several decades (Bergman, Beyth-Marom, Nachimas, Gradovitch, & Whittaker, 2008).

Group information repositories provide an online location for storing and organizing shared files where workgroup members may find and access them. Organizations and groups use these systems to store and retain important information separately from the minds and personal files of the people who contributed it (Hertzum, 1999). Advances in networking technology and applications, and falling prices for digital storage mean it is easier and cheaper than ever to maintain a group information repository (Boh, 2007). The files stored in these repositories represent explicit, codified knowledge that has been purposefully captured and managed by people; this is in contrast to tacit knowledge that has not been captured in a formal way (Cowan, David, & Foray, 2000).

These files are shared in the sense that they can be accessed by anyone with permission to use the repository; however, the action of adding a document to a repository is more like making the document available to the users of the repository, than sharing it directly with any particular person. For example, when one person emails an attachment to another person, the sender has an expectation about who will receive that file, and that the recipient will do something with it (Bellotti, Ducheneaut, Howard, Smith, & Grinter, 2005). Such expectations don't make sense when contributing to a repository—simply making a file available does not guarantee that another user will be aware that it is there, or know where to look for it. This distinction is important when one starts thinking about the repository as a medium for communication between information producer and consumer. If producers do not have a specific recipient in mind, ho do they imagine using the files they have contributed? With whom are they communicating?

Group information repositories are also different from Internet-scale repositories like Wikipedia. A repository user is generally familiar with other users through interactions that take place face-to-face in the workplace or via some communications medium, and with projects and joint work activities they are engaged in together. However, he can expect to be familiar with only some of the documents stored in a shared repository, and he may or may not have been involved with contributing and organizing documents. Also, access to group information repositories is usually restricted to a particular group of people, rather than allowing anyone in the world to obtain the information.

In a personal repository like a laptop hard drive, the information producer and consumer are necessarily the same person. However, in a situation where multiple users have access to a shared repository, this is also not necessarily true. The producer and consumer roles can be filled by the same individual, or any combination of users, resulting in a situation where a user might be trying to find documents with which she is unfamiliar or looking for familiar documents stored in unfamiliar places. Users also have different preferences for how they like to organize information, which is a problem for information management in a shared repository that does not happen in personal repositories (Berlin, Jeffries, O'Day, Paepcke, & Wharton, 1993; Whittaker & Hirschberg, 2001).

Finally, unlike library classification schemes created for describing content items and codifying relationships between subjects (Rafferty, 2001), group information repositories generally do not have rules for what the information structure should look like, nor do they necessarily have unified goals or purposes to guide users' choices about how to label and organize. And even in instances where rules exist, they are often not strictly enforced (Berlin et al., 1993; Mark & Prinz, 1997; Trigg, Blomberg, & Suchman, 1999). Having an *unstructured* organization scheme means that less effort is required when storing and labeling documents—users are free to express what is salient to them about the documents, rather than what might fit within the classification scheme (Marlow, Naaman, boyd, & Davis, 2006)—but this lack of predictability and consistency is a source of frustration for users. Repositories tend to accumulate content over time and become more and more disorganized, such that users have difficulty finding the documents they need (Rader, 2009).

1.2 The Cost of Losing Information

Working with documents is an important part of knowledge workers' daily activities, and not finding information when it is needed can be costly (Boyd, 2005). For example, in a report completed for Xerox, Inc. in 2005, IDC estimated that more than 60% of the documents

exchanged between an organization's employees are electronic[2]. Knowledge workers report spending 37% of their time on average working with documents, and 54% of this 'document' time looking for information. Half of the time, they don't find what they need (Boyd, 2005), and this can result in serious, negative consequences (Blair & Kimbrough, 2002). Lost work time alone adds up to estimated costs of $5.3 million per year for an organization employing 1,000 knowledge workers; re-creating the information costs an additional $4.5 million per year (Feldman, Duhl, Marobella, & Crawford, 2005).

Here is one extreme example of just how costly losing information can be. The US Government Accountability Office published a report in March 2009 about the National Nuclear Security Administration (NNSA) "Stockpile Life Extension Program", which is a program to refurbish stockpiled nuclear weapons that were created between 1940 and 1960. They described a problem with the refurbishment of one common type of warhead, the W76. Apparently the process requires a very particular chemical compound called "Fogbank"—and nobody can remember how to make it. according to the GAO report:

> "NNSA had lost knowledge of how to manufacture the material because it had kept few records of the process when the material was made in the 1980s and almost all staff with expertise on production had retired or left the agency" (Aloise et al., 2009, p15).

The report includes information about the costs incurred because the mysterious "Fogbank" component turned out to be exceedingly difficult to manufacture. The NNSA initially spent $22 million trying to figure out how to make the stuff again. When that effort failed, they spent an additional $23 million to invent and manufacture a substitute. And in the meantime, they spent $24 million to keep the production facility ready to go at a moment's notice if and when the problem was solved.

According to Grudin (2006), finding and reusing information is difficult for organizations. He provides an example similar to the "Fogbank" story above, in which a large pharmaceutical company found that re-testing some of their products was actually less expensive than finding the original results.

It sounds surprising on the surface that knowledge workers spend so much time looking for information, or that it could be so costly; however, as Gordon (1997) explained,

> "Why would busy, professional people spend so much time looking for missing documents? Because certain information is *mandatory* for business to be conducted effectively. If a document can't be located, it can add to the time it takes

[2]IDC findings are based on 15-min telephone surveys with 550 employees of companies from seven industries: banking/brokerage/securities/financial services, insurance, high-technology manufacturing, discrete manufacturing, energy/utilities, retail/wholesale and transportation, and other. Companies surveyed had at least 500 employees.

> to complete a task, delay its completion, or prevent it from being completed altogether A document can encode intense, sustained intellectual activity for which individuals are highly trained and well paid. Such knowledge is part of the backbone of an organization" (p112).

These statistics and anecdotes illustrate a growing problem, and organizations are increasingly turning to information systems for the solution. A recent article in the New York Times (August 7, 2009) reported that Microsoft SharePoint, the enterprise content management software suite that includes group information repository functionality, is their "hottest selling server-side product ever" and has continued to make money in 2008-09 even as the world has weathered a recession. And Global Industry Analysts, Inc. reported in April 2009 that they expect the global content management software market to reach $10.45 billion by 2015.

1.3 Purpose and Research Questions

My main goal for this research was to understand why users struggle to find the information they need in group information repositories. As I described above, these systems are different from other ways people store and share information, like email, internet-scale or personal repositories, and libraries. One important way they are different is that they are used to support the activities of multiple people working towards a common goal. However, repository systems include little functionality to support social processes. I wanted to find out whether the social dimension of these systems contributes to users' difficulty in finding files, thereby identifying or ruling out a possible avenue for future design of repository systems.

A group information repository is more complex than just an "aggregate of every individual's contribution" (Jian & Jeffres, 2006). Managing group information in a shared repository is a collaborative effort, and the information structure of the repository emerges through users' individual, idiosyncratic labeling and organizing choices. Labeling (what to call a file) and organizing (where to put a file) in a group information repository can be considered an act of *packaging* files for later reuse; unfortunately, in most situations, people do not package content effectively for reuse by others (Markus, 2001). Coordination in real-time communication depends upon a back-and-forth exchange between conversation partners who monitor each other for signs that they are being understood (Clark & Brennan, 1991; Clark & Krych, 2004); however, the transmission of any sort of social feedback is not possible in most group information repository systems.

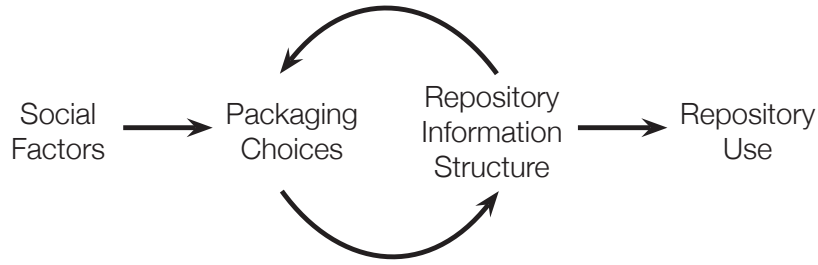


Figure 1.1 Diagram depicting the phenomena explored in this thesis.

In this thesis, I focus on *packaging* as a social process with important consequences for sharing information online. At a high level, my hypothesis is that users of these systems make choices about how documents should be labeled and where they should be stored in relation to other files in the repository, and their choices are influenced by knowledge, beliefs and assumptions about other users. These choices determine how the information in the repository is structured, and the information structure affects whether or not users can find what they need. The information structure is co-constructed and evolves over time; early choices of individual users can constrain later choices and thus the ability of others to successfully find information. Figure 1.1 depicts these relationships at a high level; I will revisit this diagram in the final chapter.

I asked the following specific research questions:

1. How do users share information using group information repositories? What influences their packaging choices, and how do they manage the information in the repository?

2. How do common ground and audience design affect file and folder labeling and organizing choices? How do these choices affect subsequent finding behavior?

I conducted two studies to answer these questions. First, I interviewed users of a real-world information sharing system in order to better understand how packaging choices are simultaneously constrained by social concerns, and also influence others' future information management tasks and behaviors. This study also allowed me to collect data reflecting group-level phenomena, in a setting with high external validity. I then conducted a two-phase experiment; in the Organizing phase I investigated whether communication processes might play a role in shaping packaging choices when organizing and labeling files for others. In the Finding phase of the experiment I measured the impact of different packaging choices on search task behavior. This study allowed me to explain how individual-level processes

influence information structures, and connect finding behavior with the conditions under which the structures were created.

These separate investigations allowed me to study aspects of information sharing systems in two contexts, using a multiple method approach (Creswell, 2003). By using both qualitative and quantitative methods, I am able to start to fill in some of the gaps in understanding due to limitations in the different methods (Tashakkori & Creswell, 2007). The interview study allows me to learn about user behavior in a real world context, but it cannot produce causal claims or generalizable results. The experimental study can address some of these shortcomings, but at the expense of external validity. My main objectives in this research were to gain new knowledge that would both contribute to the growing literature on group information management (GIM, in contrast to personal information management or PIM), and also serve as a first step toward creating better information sharing systems that will help people find the information they need when they need it. To that end, the different views into the phenomena of packaging information for future reuse are integrated in this thesis to present a more detailed picture of the social influences present in information sharing systems.

1.4 Thesis Outline

This thesis is organized into seven chapters, including the Introduction. The next chapter presents a literature review that describes "packaging" for future reuse as a social process, and provides background information both about what is already known about digital information management and sharing, and relevant psychological theory.

In Chapter Three, I present the method and findings of the interview study, and describe implications for the design of group information repository systems that grew out of the findings. Chapters Four, Five and Six all describe aspects of the experiment; Chapter Four contains a description of the method and procedure, and a discussion of the research goals and experiment design choices that went into creating the procedure. Limitations of the design are also introduced. The method and procedure are followed by the Finding phase results in Chapter Five; although the Organizing phase took place first, the Finding phase results are of primary interest. The results of the Organizing phase, presented in Chapter Six, are included to help tease apart what caused the interesting patterns in the data from the Finding phase.

Finally, in Chapter Seven I summarize and integrate findings from the interview study and experiment, and revisit the discussion of limitations started in Chapter Four. I also

present implications for both theory and design, and suggest both immediate next steps and directions for future research. I include several Appendices at the end of the thesis, including the interview protocol, complete instructions and materials for both phases of the experiment, additional statistics and complete regression output, and the R code (R Development Core Team, 2009) I wrote to help with parts of the analysis.

Chapter 2

Literature Review

2.1 Introduction

Group information management behaviors and systems are often studied in terms of an individual's efforts toward organizing and using information, rather than focusing on the aggregate or group level effects (Lutters, Ackerman, & Zhou, 2007). For example, Voida et al. (2006) created the "Sharing Palette" for granting other individuals access to one's personal files that makes information about what was shared with whom more visually explicit. Whalen et al. (Whalen, Toms, & Blustein, 2008) designed a "File Manager" and "Sharing Console" to help users become more aware of the usage history of their shared files. And, Tang et al.(Tang et al., 2007) built "LiveWire", a system that is able to detect similarities and differences among individual enterprise knowledge workers' files; they suggested this kind information could be used to help other individuals find information they need. These approaches are similar in that they focus on providing users with more information about the objects in the system—where they are, who is looking at them, how they've been used in the past, etc.

In this literature review I argue that files in a group information repository are "shared", and this means repositories are instances of "social" systems in the same way that other technologies for sending and receiving files are "social". I appropriate the term *packaging* from the knowledge management literature and use it to refer to organizing and labeling information that is contributed to repositories. I then review what previous research has learned about how knowledge workers organize and label files, both for themselves and in group information systems.

Packaging is a social activity, in that the packaging choices of one user can affect the reuse of that information by another; I suggest that theory from communications and psychology can be used to help us understand how and why people make the packaging choices

they do. I provide an overview of how common ground is usually studied in the lab, because this literature informed the design of the experiment described in Chapters 4 through 6. Finally, I discuss how people go about finding information in group information repositories, which depends on users recognizing the information they are looking for based on the labels and folder groupings created by others.

2.2 "Shared" Information

Users of group information repositories can be categorized by the role they play with respect to the information in the repository. Some users are "producers," file authors who create content and contribute it to the repository. Others are "consumers" who are primarily re-users of the information they retrieve from the repository (Markus, 2001). In a personal repository like a laptop hard drive, the producer and consumer are the same person. However, in a group information repository this is not necessarily true—these roles can be filled by any combination of group members who have access to the "shared" files in the repository.

What, then, does it mean for a file to be "shared" in this way? The Merriam-Webster online dictionary (http://www.m-w.com/) defines the verb "to share" as "to have in common" and "to tell to others". There are of course many ways for people to share files. Often when people say they are "sharing" a file electronically they mean in the "tell to others" sense, sending and receiving via email, which is the most prevalent method used in organizations (Voida et al., 2006). But email is not the only way to share files. Group information repositories are another, as are CD's and DVD's, instant messaging, hard copies, websites, MMS messages from mobile phones... this list continues to grow. These various methods can be classified according to three dimensions:

- *push-pull*: push is explicit sending from one person to another, initiated by the producer, vs. pulling information from a website or library, initiated by the consumer
- *physical-digital*: sharing "physical" files by handing over printed copies or CD's, vs. emailing documents or viewing web pages
- *controlled-open*: controlled means limiting access to specific people, for example by using permissions and some authentication method, vs. open access where anyone can view or retrieve the information

Figure 2.1 illustrates the relationships among several different methods used to share files in different forms (i.e. a book might be considered a hard copy of a file). For example, email and instant messaging are two methods for sharing files that correspond to the push,

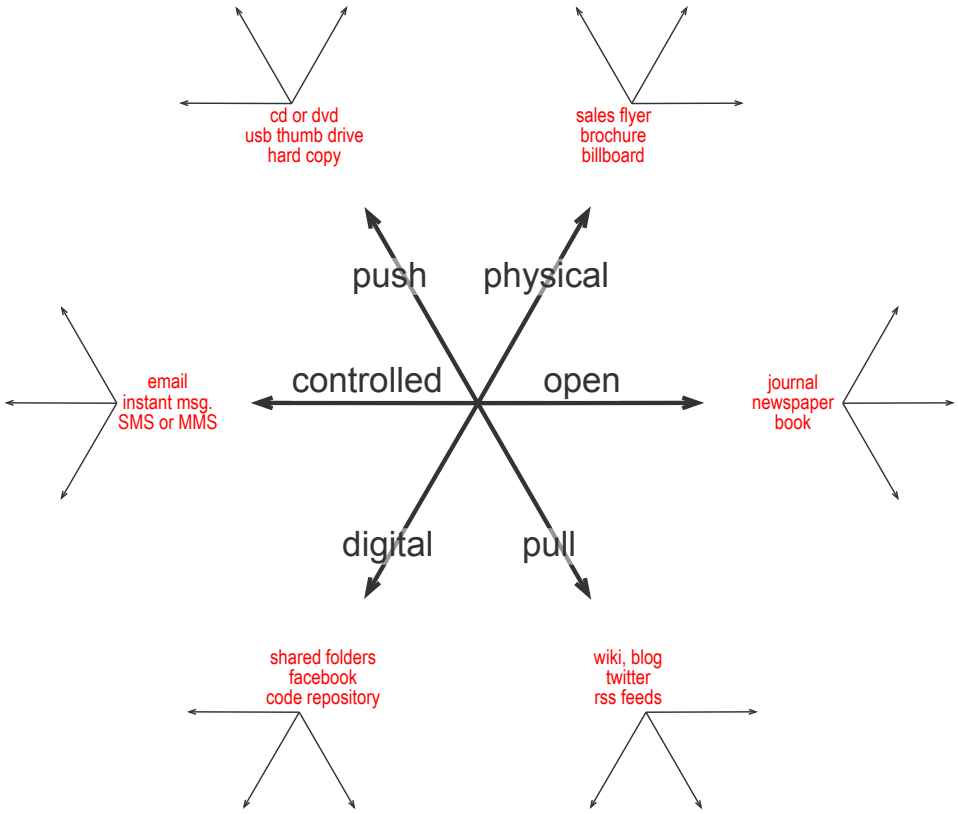


Figure 2.1 Classification of file sharing methods on three dimensions

controlled, and digital dimensions in the figure. Via email, digital files are pushed by the sender, or producer, to a specific recipient. In contrast, books and periodicals in a public library correspond to the pull, uncontrolled, and physical dimensions. They are retrieved by the consumer, freely available, and in hard copy form.

I suggest that sharing files can also be accomplished by making those files accessible to a specified group of people at all times, as with a group information repository (pull, controlled, digital in Figure 2.1). Sharing in this way involves separate actions by the producer and consumer, potentially with no intended recipient or use specified. This is a slight change to the typical conceptualization of what it means to "share" a file electronically – availability and access as opposed to transmission and possession. However, this still may be considered sharing in the "have in common" sense.

The distinction between producers and consumers it is an important one: it implies that there is an exchange of information that occurs between users who are producing content, and users who are accessing content, via the repository. This information exchange includes not only the files themselves, but also the information contained in the document and folder

labels and the hierarchy structure. This makes group information repositories different from other methods for sharing files, and therefore assumptions and findings about user behavior file sharing in other contexts are not be directly applicable to group information repositories.

2.3 Packaging: Organizing and Labeling Files

When storing a document in a group information repository, information producers must decide where the document will be stored, i.e., the folder location in the hierarchy at which someone else will be able to access the document again, and give the file a textual label. Markus (2001) referred to packaging as, "the process of culling, cleaning and polishing, structuring, formatting, or indexing documents against a classification scheme" (p. 60), preparing them for future reuse. This is exactly what repository users are doing when organizing and labeling files. Hertzum and Pejtersen (2000) wrote about two case studies of engineers' information seeking, focusing on finding information in documents. Regarding packaging, Hertzum and Pejtersen wrote:

> "Packaging also requires that the [engineering] professionals suspend their normal way of looking at and working with their documents to take an outsider's look at them. This is, however, difficult because the individual professional has an inherently incomplete sense of whether his/her documents will eventually be of interest to someone else, and, if so, to whom and in what context." (p. 47).

In other words, simply being aware of others' knowledge, background and joint experiences is insufficient for properly "packaging" information for a group information repository. The ability to take the perspective of others is also necessary. Interestingly, this problem does not occur exclusively in group information repositories. It even occurs between professional catalogers and information seekers. Šauperl (2004) interviewed 12 catalogers about their process for cataloging, and concluded that they were more concerned about common ground with other catalogers than with people who might be using the catalog entries they were creating. There are at least three possible perspectives from which the meaning of any given document may be interpreted: the author's, the cataloger's, and the reader's. Šauperl found that the catalogers who participated in the study were aware of this, but mainly tried to stick to the ways similar content had been cataloged by other catalogers in the past, rather than anticipating potential readers' perspectives. According to Šauperl, this seemed to be inherent to the indexing process which requires adherence to structured formats, and that consistency be maintained with the way similar content items have been cataloged in the past.

Once an information consumer has decided to look for information in a repository, he must browse the repository and make judgments about which documents might contain the needed information. These judgments are difficult in group information repositories, because contextual information essential to understanding and interpreting the information in the repository is typically not captured in the information structure (Hertzum, 1999). While a project is active, this may not be much of a problem, because those involved are familiar with the context. But once the project is over that knowledge is rapidly lost (Hertzum & Pejtersen, 2000).

2.3.1 Categorization

Categorization is a cognitive shortcut that allows us to predict, infer, and assume facts and relationships among things we encounter in our daily lives. Psychologists have been able to uncover many aspects of mental categories over the years, but there is still no consensus about the underlying cognitive mechanisms that allow us to form categories (Murphy & Lassaline, 1997). Categories to some extent reflect regularities or structure in the nature of what's being categorized; they are not arbitrary (Mervis & Rosch, 1981).

Inductions from categories are a fundamental process of human cognition. We automatically assume that things grouped together in a category have something in common, even if that commonality is not immediately observable. This implies that there's a learning component to categories—by encountering categories others have created, like in a group information repository, and trying to infer what a set of files has in common, we could actually be learning about how others think about the information (Gelman & Markman, 1986)

People prefer to use a so-called "basic level" when naming categories—the basic level is specific enough to distinguish members of the category from other basic level categories (i.e., cat vs. dog), but not so specific that one must bring to mind specifics that are unnecessary for telling the categories apart (i.e., poodle vs. siamese). The basic level of categorization is cognitively efficient because it maximizes within-category similarity relative to between-category similarity. According to Bates (1998), there has not been an investigation into how well formal classification systems correspond with basic level categories. This comparison has not yet been undertaken for folders in personal or group information repositories, either.

Jacob (1995) wrote, "category membership is not contingent upon a set of shared attributes or properties but is determined, instead, by the individual's recognition of integral relationships that exist between the observable properties exhibited by a set of category members." Folders in a group information repository are essentially categories, created

based on perceived relationships among the information in the files. However, these relationships and users' perceptions of them are not stable over time, even within the same information producer. In addition, 'typical' files for which there is an obvious category are easy to assign to a category; but, for atypical files that could exist in multiple different categories, people tend not to agree on where they should go because they choose different features upon which to base their category assignments (Mervis & Rosch, 1981). Finally, categorization differences exist between experts and novices in a domain (Kellogg & Breen, 1987; Marchionini, 1997; Murphy & Lassaline, 1997). Expertise might cause differences in the perceptions of the relationships among the files in a group information repository, and also how these relationships are represented by the categories to which they are assigned. These inter-user differences could make it difficult for experts and novices to use each other's file structures.

There is one other difference that comes to mind when filing or categorizing in a group setting. Suchman (1994) cautioned that categorization serves not only to make things more organized; it can also communicate information about the values of a group, and in essence be a form of social control. Document labels and the representation of the relationships between content items and people that are made explicit in a hierarchy structure can clearly communicate what, and who, are "important" and what is not, and reflect power structures within the group.

Not all members of a category are equally representative of a category; "gradients of representativeness" exist, and people are generally able to judge whether a particular exemplar is typical or atypical of a category (Mervis & Rosch, 1981). The catch is that these judgments tend not to be consistent for most categories, both within and between subjects. In a group information repository, if a particular file is not a very good match with any of the existing folders (categories), one might expect that it could end up in one of many different folders, and there would be a lot of variability in the locations information producers would choose. This inconsistency in category assignment would make it unlikely that an information consumer would look for the file in the correct place.

2.3.2 Organizing

A few papers have been published that focus on how people organize information in their physical and digital workspaces, mostly in the Human Computer Interaction (HCI) literature. Researchers studying how people organize information in their physical and digital workspaces commonly conduct semi-structured interviews in which they ask participants to describe how the documents and other information in their offices or on their computers

is organized. A small number of papers have also reported analyses of participants' folder structures for clues about the rules or strategies that went into creating them, or solicit survey responses in order to gather quantitative data.

One of the most well-known findings in personal information management (PIM) came out of a workspace organization study by Malone (1983): some people are "filers" and others are "pilers". Malone interviewed professional and clerical workers, and found that some had elaborate file systems for categorizing their information, while others left things laying around in piles around the office. In a study of email organization behavior, Whittaker and Sidner (1996) found that peoples strategies for dealing with their incoming mail fell into three categories, "no filers", "frequent filers", or "spring cleaners" (people who periodically engage in 'cleaning up' their email folders). Another finding of these studies had to do with deliberately storing items in certain visible places so they would serve as reminders.

Hertzum (1999) wrote about how documents move from "action to archive", from "spatial, loosely systematised, memory-based organization to category structures" (p. 43) as they become less and less important to daily work activities. Files that are frequently used tend to be stored in places that are quickly and easily accessible; older information that is used less often is organized into more concrete and complex categories that are less visible (Barreau, 1995; Malone, 1983).

Analyses focusing on salient features of the content and context of files describe the relationships among these features, and their relative importance. The list features researchers have identified is quite long, and includes things like frequency and recency of use, salient aspects of the situation surrounding the document at the time it was filed (Kwasnik, 1989). Barreau (1995) summarized findings from studies of personal electronic file organization they had each conducted separately a couple of years earlier. They described three types of information among users' personal files: ephemeral (changes often, like a to-do list), working (used frequently), and archived (used infrequently). It is conceivable that group information repositories might at least contain working and archived information; it is less clear whether repositories would be used to store ephemeral information. In a later study of personal information management, Boardman and Sasse (2004) found that most files their participants mentioned retrieving fell in the "working" category. They also reported that participants sometimes mentioned accessing older items as well, "We found that although older items may be accessed erratically, they can be highly valued by people" (p. 587).

When engaged in "document triage"—i.e., trying to decide how to organize a new or incoming file (Bae et al., 2006), there are three high-level problems a person faces (Whittaker & Hirschberg, 2001):

- figuring out the "value" of incoming information, whether it is important or needed

- figuring out how to categorize the information
- deciding where to put the information, or deferring judgment

Whittaker and Sidner (1996) wrote that organizing files is a "cognitively difficult task", because when an information producer is deciding where to put a file, he must imagine where he and others might want to go looking for the file again, as well as remembering how everything else is categorized, the rules and definitions for what each folder contains, and the relationships among the different folders. The consequence for making a wrong choice is that nobody will be able to find the file again. Making this choice gets harder as the repository gets larger, because it is not possible keep all the folders and all the rules in one's head at the same time (Bellotti et al., 2005; Malone, 1983; Whittaker & Hirschberg, 2001). The more folders one has, the less helpful they are at reducing the information one has to remember about where files are located, because instead one has to remember all the different rules for what the folders represent. If each folder contains two or three documents, there are a lot more folders to remember than if each one contains ten or fifteen documents.

At the outset of using a repository, users don't know what information structures will work best (Barreau, 1995). The information structure in a group information repository is resistant to change, and does not allow for back-and-forth interaction between the producer and consumer. The structure evolves slowly over time, as files are incrementally added. Bae et al. (2006) wrote, "...people create and refine categories incrementally as they read documents and decide what to do with them. Thus the categories that have already been defined influence subsequent searching and reading just as reading affects searching and organizing."

2.3.3 Re-Organizing

Filtering and pruning are activities that information producers typically don't like to do (Markus, 2001), and increases in digital storage space mean that people are able to store more information than ever before. So they defer evaluation, or initially put aside documents that are hard to classify, and only deal with them later if something else happens to prompt action. If this doesn't happen fairly quickly after the file is put aside, it probably won't happen at all (Whittaker & Hirschberg, 2001). Boardman and Sasse (2004) reported that the knowledge workers who participanted in their study did not engage in maintenance or updating of their folder structures, except during times of major change such as starting a new job.

These findings came from a study of personal information management, but there is no reason to suspect that they would be invalid for group information repositories. In fact,

users of a shared repository might be even more reluctant to purge. Imagine a refrigerator in a common area in a workplace. Food accumulates in the refrigerator over time as people forget what they've brought or it gets buried underneath the new arrivals. The older food starts to go bad and get moldy. Eventually, someone just gets disgusted and fed up and starts throwing things away. A similar phenomenon can also occur in other workplace common areas — consider the example of a hoteling office. When several people share an office on a temporary basis, stuff can accumulate just like it does in the refrigerator. However, it is much harder for one person to make an executive decision to throw away other people's stuff when there are no outward signals of spoilage to indicate that the items will no longer be needed. Clearly, nobody will want the moldy pizza; but the choice may not be so black-and-white for a pile of old meeting minutes or out-of-date lab procedures. The occupants of the office must then communicate about the task of cleaning up the office, in person if they happen to run into each other, but more often by leaving notes in the office.

In a group information repository this problem is compounded further because there are not necessarily cues in the interface indicating how recently an item was accessed, and the costs of leaving outdated digital information 'laying around' aren't as immediate as having to push someone's pile of books and papers out of the way to set up your laptop in the hoteling office. In addition, mechanisms rarely exist within a shared repository interface to communicate about a specific file or folder; these communications must be conducted in another software program, or another medium altogether. The path of least resistance is to leave things as they are. In other words, once a repository has been organized, it is very difficult to change the structure to suit individuals' information needs (Boh, 2007). Cleaning up a repository is too onerous a task for most users to be willing to undertake (Barreau, 1995).

2.3.4 Labeling

Conventions are spoken or unspoken rules for how people should behave in certain social situations. Such rules, even in distributed collaborative systems, evolve as the system is used (Ackerman, 2000; Krauss & Fussell, 1991). With respect to group information repositories, conventions are typically considered to be rules for how files should be labeled, i.e. "naming conventions". While users tend to have their own personal mental rules for how their files are labeled (Carroll, 1982), relying on joint conventions shared by a group to keep a repository organized is rarely successful without significant overhead such as incentives or strict enforcement (Berlin et al., 1993; Mark & Prinz, 1997; Markus, 2001).

Berlin et al. (1993) encountered problems with conventions when they implemented

their own "group memory" system for their research group. Despite agreeing upon how files should be labeled in their repository, there were differences in how group members adhered to them. One member of the group commented, "It was hard to remember what we'd agreed to, and what each person remembered tended to drift toward the person's initial position." (p. 26).

Sometimes, conventions are agreed to in principle, and then intentionally ignored in practice. Mark and Prinz (1997) conducted a field study of a group using a "large groupware system" to store and share documents. The users of this system held "workshops" in the early days of using the system in order to discuss and decide upon conventions they would follow. After using the system for about six months, it became clear that it was becoming disorganized and unusable, in part because no one was adhering to the conventions. Mark and Prinz (1997) observed that "different users prefer to store and access documents differently." She concluded that in this case, it had been too difficult to imagine in advance what conventions would be needed. As the system was used, work practices changed, making the conventions the group had agreed to less appropriate for the situations that arose. Also, there were some users who were unwilling to give up their own, idiosyncratic practices. In some cases it was a conscious choice to violate conventions. One user said, "Naming conventions, reference code, and subject area, I always violate. I give file names that seem to fit" (Mark & Prinz, 1997, p. 23).

One hurdle for naming conventions is a robust property of label choices first identified by G. Furnas, Landauer, Gomez, and Dumais (1983). They reported that random pairs of people use the same label for an object at most 20% of the time. They explained this phenomenon by writing: "There are many names possible for any object, many ways to say the same thing about it, and many different things to say. Any one person thinks of only one or a few of the possibilities" (p. 1796). G. Furnas et al. completed several studies of labeling in a variety of contexts: "the verbs used in spontaneous descriptions of the operations needed to perform manual text-editing operations, descriptions of named common objects designed to induce another person or a computer to return the name, superordinate category names for items available in a swap-and-sale listing similar to classified ads in newspapers, and index words provided for a set of main-course cooking recipes." The percent agreement ranged from 7% to 18% in these studies (G. W. Furnas, Landauer, Gomez, & Dumais, 1987).

Many other researchers (e.g., Bates, 1998; Trigg et al., 1999) have also observed the same pattern, referred to by G. Furnas et al. (1983) as the "vocabulary problem". The implications of these findings for group information repositories are dire: if left to their own devices, people are extremely unlikely to use choose the same label for the same file, and it takes considerable effort to overcome this tendency (Greenberg, Crystal, Robertson, &

Leadem, 2003).

2.4 Common Ground and Audience Design

Common ground is the mutual knowledge, beliefs and assumptions that people share about each other (Clark & Brennan, 1991). Humans' use of language is imprecise and flexible, and meaning is determined by the surrounding context and complex communication processes. As a conversation progresses, participants introduce ideas and vocabulary that become part of their common ground, and can subsequently be referred to without the overhead of having to re-introduce them. Common ground is necessary for coordination of conversation, and essential for people to understand one another.

Conversation participants believe common ground exists when there is evidence for a "shared basis". Evidence that a shared basis exists for members of a workgroup using a shared repository can be recognized in the usage of specialized knowledge and language (Clark, 1996). Clark also wrote that conversation participants develop a "feeling of others' knowing" (p. 111), a sense of what others do or do not know, that plays a role in assessing how much common ground exists between them. Common ground can be classified into three types (Nickerson, 1999):

- *Shared immediate context* which is ephemeral, existing in the present while two people are in a conversation or working on a task together.
- *Shared past experience*, which is delineated by delineated by contemporaneous and collocated past interactions and experiences; i.e., people who have interacted with each other in the past. This type of common ground is created among people who might have taken the same class at the same time and experienced the same events, or worked together on a group project.
- *Community or category membership*, shared by people who have characteristics in common but have never directly interacted. For example, two people who both grew up in Chicago but never met can be said to have community membership common ground. Likewise, two experimental psychologists share this type of common ground, where an experimental psychologist and an accountant would not.

There is some debate in the literature regarding what the cognitive representation of common ground might look like. Clark (1996) does not address the issue of mental representation of common ground; others in the literature have attempted to clarify how a seemingly endless reflexive process might be represented; Lee (2001) calls this the "*mutual knowledge*

paradox". According to Nickerson (1999), common ground can be conceptualized as a "model of others' knowledge". He argues that "one's behavior with respect to others is influenced in various ways by what one knows (i.e., believes, assumes) about what specific others know" (p739). Conceptualizing common ground in this way makes sense from the standpoint of an applied researcher, like me, who is more interested in the implications of common ground than the cognitive processes by which it is formed, represented, and used.

Interlocutors develop a mental model of what they assume other individuals know and expect, and use this information to effectively tailor their utterances to their audience (Nickerson, 1999)—this activity is referred to in the literature as *audience design*. For example, Schober and Clark (1989) demonstrated the effects of audience design in an experiment using a twist on the canonical referential communication task (see below for more information about these tasks and how they are used to study common ground). In the experiment, one participant instructed another how to construct an abstract shape using puzzle pieces. A third participant (the "overhearer") who was not visible to the others and did not speak during the experiment listened in and tried to construct the same abstract shape with another set containing the same pieces, at the same time. The intended listener, or "target audience", was significantly more accurate at constructing the shapes than the overhearer (98% to 85%).

Beliefs about the goals of the listener also affect how speakers construct their utterances. A. W. Russell and Schober (1999) found that being correctly informed about a partner's goals had an impact on how much was said and how understanding was displayed. Also, participants who were given no information about their partner's goals assumed others shared their goals by default.

The two experiments discussed above involved synchronous conversation. An experiment conducted by Fussell and Krauss (1989) showed that people label things differently for themselves than for an unknown future person. Participants wrote short descriptions of abstract line drawings to help themselves identify the drawings at a later time, or to help someone else identify them. Descriptions were more than twice as long when written for others than for themselves (12.7 versus 5.0 words). When participants returned weeks later, they used the descriptions to identify the drawings. They were correct 86% of the time with their own descriptions, 60% of the time with descriptions written for others, and 49% of the time with descriptions written by other people for themselves. Subjects also had the highest confidence that they had identified the correct shape based on their own descriptions, followed by descriptions written for others, and finally descriptions by others for themselves.

The results of these experiments indicate that common ground might indeed affect the labels information producers create for documents they store in a shared repository.

People tailor what they say to whomever is the intended recipient, even when they are simply instructed to write descriptions for "someone else". While a shared repository is not a communications system, language is being used in the form of labels to represent the contents of documents, and also to suggest relationships among groups of documents. Common ground helps us understand each other in conversation; the same might be true when the communication is mediated by a shared repository. Groups with more common ground might assign labels to documents that others in the group will be able to anticipate more frequently than 20% of the time. However, an important difference between a real-time conversation and any type of asynchronous communication is in the timing of feedback, which is essential for establishing common ground and negotiating meaning (Clark & Brennan, 1991). For example, facial expressions are a form of nonverbal feedback that convey whether or not a speaker has been understood by a listener.

2.5 Referential Communication Paradigm

Referential communication tasks are commonly used in psychology experiments investigating common ground and other aspects of language use in conversation. Researchers often prefer to study common ground in the lab, rather than in the field; as Schober and Brennan (2003) wrote:

> "Laboratory studies of task-oriented conversation have the advantage of allowing researchers to assess speakers' intentions and adressees' comprehension independently of the conversation, through external behaviors like grasping and moving objects." (p129).

In a typical referential communication experiment, two conversation partners must work together to complete some kind of task. One partner has information the other does not, and they must talk with each other to successfully complete the task. In some instances, the partners are in the same room but not allowed to see each other; in others, communication occurs via a medium such as video or instant messaging. Other variations included using pre-recorded speech in lieu of one of the partners (Chantraine & Hupet, 1994), or an overhearer who tried to complete the task without directly interacting with the conversation partners (Schober & Clark, 1989), or even a confederate as one of the partners (Metzing & Brennan, 2003). Referential communication tasks are often measured in two ways, either by counts of words and speaking turns, representing 'efficiency', or by evaluating the task outcomes as a measure of effectiveness.

According to Carroll (1980), the three phases that occur in a referential communication task are:

1. trade descriptive phrases (for the object, image, abstract form, etc.)
2. label proposed by one person
3. label accepted by other person and both use it henceforth

A group information repository is a kind of "common workspace" within which all users see basically the same information, although not necessarily at the same time (Gergle, Kraut, & Fussell, 2006). The descriptive phrases, or "referring expressions" in a referential communication task are like the file and folder labels in a group information repository. Users of group information repositories and engage in activities and tasks which require that they find documents via these labels in order to use them. However, there are two crucial differences between the repositories and typical referential communication tasks (Brennan & Clark, 1996; Fussell & Krauss, 1989; Metzing & Brennan, 2003):

- Partners in a referential communication task repeatedly interact with one another, and develop a mental representation or model model of both the history of the interaction and the other person, even when their only experience with each other is for the duration of the experiment.
- Feedback between communication partners in a referential communication task is essential for label agreement to occur. When the interaction history does not exist or the exchange of feedback is prevented from taking place, evidence that an agreement has been reached is not available. This results in communications that are less efficient, and in which miscommunications or mistakes can occur.

Referential communication tasks as they are studied in the lab are tightly coupled (Olson & Olson, 2000); organizing and finding in group information repositories are loosely coupled activities. However, a few researchers have attempted to study both categorization and common ground using the referential communication task paradigm. For example, Markman and Makin (1998) asked participants to collaboratively build fairly complex LEGO models, and then participate in a sorting (categorizing) task. One objective of this study was to find out whether communicating about a set of objects made participants more likely to have similar categories for those objects. This hypothesis was confirmed; Markman and Makin found that dyads with shared past experience common ground categorized LEGO pieces more similarly than random pairs of people. In addition, dyads who had constructed similar types of LEGO models (cars vs. spaceships) produced categorizations that were more similar to other dyads who had constructed the same type of model.

2.6 Recognition and Finding

Lansdale (1988) suggested that personal information management applications for computers should take advantage of the way human memory works, rather than mimicking the ways people manage information in the physical world. Memories are formed as people interpret meaning in a particular context, and the ability to recall details depends on the relationship between how those details are stored in memory, and what is salient about the context in which the person is trying to remember the details. Many people can identify with the experience of losing an object like one's keys, and while looking for them thinking, "Now what was I doing the last time I had them?" Not only do we remember specifics about the sought-after object; we also remember other information about the surrounding context in which it was used. In other words, it is both what we're thinking about when we store something, and what we're thinking about when we're trying to find it, that interact to determine whether or not we'll be able to achieve success.

2.6.1 Orienteering

In a study conducted by Boardman and Sasse (2004) users looking for information their personal repositories used a combination of browsing and sorting of folders. Because they were searching their personal repositories, they exhibited a tendency to know approximately where in the hierarchy to start looking. From there they used recognition memory navigate to the particular document they wanted. Teevan, Alvarado, Ackerman, and Karger (2004) called this *orienteering*: using recall to make an initial jump to a location from which to start navigating in steps, via recognition, toward the ultimate goal. At each stage, the local context is used to remind people about where they should go for the next step. Teevan et al. (2004) mentioned one participant who tried to find something in her personal repository, but could not explicitly recall the path or any of the folder labels for where it was stored, making it very difficult for her to search for the document using a query interface. Orienteering allowed the participant to find the document, because the information she needed at each step to prompt her next step via recognition was built into the information structure. All she had to do was be able to recognize the next step, not recall it.

However, one difference between shared and personal repositories is the variance among users in their level of familiarity with the documents and folders in the repository, and the vocabulary used to label them. S.-J. Chang and Rice (1993) wrote that goals for browsing can be vague and changing, and users might not know exactly what they are looking for until they see it. It might be that common ground could play a role in how browsing goals

are specified; however, it is not clear from the recognition or label-following literatures how users might consider their model of others' knowledge when cognitively matching a vaguely specified goal state with the shared repository structure they encounter, in order to make choices regarding how to proceed.

2.6.2 Label-Following

Recognition memory is triggered by some kind of stimulus or other information in the environment. There are two types of recognition, *familiarity* and *recollection*. According to Yonelinas (2002), familiarity is an automatic, perceptual process — you feel like you've seen something before, but can't remember where; recollection happens when you recognize something you've seen before, and are able to elaborate on that memory once it's been triggered. It is difficult to find studies of the implications of recognition memory processes in real-world situations, rather than lab experiments with little external validity (Elsweiler, Ruthven, & Jones, 2007); The label-following literature in human-computer interaction is one example of research with external validity.

Label-following occurs when users attempt to complete a task using a menu-based interface. Researchers studying label-following were inspired by the cognitive psychology problem-solving literature. The array of choices among possible menu items is the *problem space*, and novice users match the vocabulary they see in the task description with the labels they can see in the interface when choosing what to do next (Polson & Lewis, 1990; Franzke, 1995). A label-following perspective was incorporated into the cognitive walkthrough usability inspection method question, "Will the user notice that the correct action is available?" (Wharton, Rieman, Lewis, & Polson, 1994).

In one label-following experiment, Mehlenbacher, Duffy, and Palmer (1989) hypothesized that user tasks and goals might affect label-following in menus. In the first condition, the task description contained the same vocabulary as the menu items (direct match); in the second condition the task descriptions contained synonyms of the menu vocabulary (synonym); and in the third condition they used pictures and diagrams to communicate what the users were supposed to be trying to do with the software (iconic). Unfortunately, Mehlenbacher et al. (1989) did not ask users in the iconic condition to verbalize their interpretation of the pictorial task descriptions. The menu in this study was not hierarchical; it was a flat list, and menu item labels were grouped either alphabetically or functionally. They found that performance was fastest and nearly error-free for the 'direct match' task combined with the alphabetic menu. The functional menu structure was faster for the synonym and iconic task types. There were more errors in the iconic tasks than the synonym tasks; even

when they analyzed just trials without errors (for the iconic tasks only), the functional menu was still faster than the alphabetic menu. All conditions showed practice effects such that difference between conditions were nearly eliminated after the first few trials. The results of the Mehlenbacher et al. (1989) experiment indicate that when users are left to generate their own goal specifications, more variability exists in measurable performance outcomes of their label-following behaviors.

2.6.3 Browsing

In physical space, people make inferences and assumptions about where things "should be" located based on information in the environment. For example, everybody has had the experience of looking for the bathroom in an unfamiliar building — there are places where you just expect to find a bathroom, based on your past experience in other buildings and cues from what you see around you. Information spaces that are arranged in a hierarchical structure have built-in explicit cues about what is located where. Hierarchies may convey information about the structure and content of a shared repository that information consumers would be unable to access if they were to interact with the repository using a search interface only. According to Dourish (2004), "In information work, the meaningfulness of information for people's work is often encoded in the structures by which that information is organized" (p. 30). Jones, Phuwanartnurak, Gill, and Bruce (2005) found that folder hierarchies and document labels provide meaningful information that helps people summarize content as well as organize it. Grouping things manually allows for the formation of visible relationships between documents. Visibility into the relationships in an information space might allow an information consumer to orient herself to the content, and choose better where to go next (Chalmers, 2003). It is possible for structure to be inferred from a list of search results and memory for the query that was entered, but this forces the information consumer to work harder to construct structural relationships that can be explicitly stated with a hierarchy (Cutrell, Robbins, Dumais, & Sarin, 2006).

S.-J. Chang and Rice (1993) describe browsing as "recognition-based" and "searching without specifying [query terms]". Information behavior researchers refer to three levels of goal-orientation in browsing (S.-J. Chang & Rice, 1993):

1. Search or directed/goal oriented browsing, i.e. "I'm looking for a specific document and I know it is here somewhere."

2. General purpose, semi-directed browsing, i.e. "I need documents related to a certain project or created by particular person, and I think they might be here."

3. Serendipity, undirected, random, not goal-oriented browsing, i.e. "I need to find all the information I can related to a past project that I was not involved with."

When browsing a hierarchical structure, how do people decide where to look next, and when to give up and move on? Pirolli (2005) wrote about information foraging theory, which accounts for and predicts browsing behavior on the web. A shared repository's information structure is similar in some ways to a website with a link structure: folder labels are like link text. An information consumer is able to browse until she recognizes something related to what she is looking for (Bruce, Jones, & Dumais, 2004; Trigg et al., 1999).

Information foraging theory states that the links on web pages are "cues" that activate certain cognitive structures related to those cues, via spreading activation. Users will choose to follow links with text that triggers higher activation levels in memory for concepts related to the user's goal state. Users move on from a given location when the expected potential of the current site (estimated from activation triggered by visible links) is less than that of moving on (estimated from past web surfing experiences). A study by Mobrand and Spyridakis (2007) demonstrated that "navigational link phrasing" — link labeling — affected navigation in a news website; confusing or ambiguous hyperlinks decreased overall comprehension of the information, and discouraged exploration. An experiment conducted by Vaughan and Dillon (2006) found similar results; they created two versions of the same health information website, one which was similar in design and link labeling to typical health information websites, and one which violated users' expectations for page layout and link labeling. The group that used the "expectation-conforming" version of the website explores more of the site initially than the group using the "expectation-violating" version, but when asked to search for information they were able to find what they were looking for faster.

2.6.4 Browsing vs. Search

Specifying a search query requires recall, but browsing a hierarchy depends on recognition, and has been referred to as "searching without specifying" (S.-J. Chang & Rice, 1993). Browsing involves visual scanning of a resource or structure, and movement through an information space, as opposed to evaluation of query results. Two studies from 2008 independently concluded that when it comes to their personal information, users prefer browsing to search; these findings are likely to apply to group information repositories as well, which are more similar in scale to personal repositories than the Internet.

Civan, Jones, Klasnja, and Bruce (2008) compared Hotmail folders with Gmail labels for organizing email—essentially a "folders vs. tags" study in the context of email. Nine

out of 10 participants in their study exclusively used browsing when asked to retrieve documents they had organized. Bergman et al. (2008), in an extensive and well-executed study, investigated the information management practices of desktop computer users to find out whether advances in search technology have made users more willing to search to find personal files. The study incorporated two search engine versions, a baseline and a more advanced engine, on both MacOS (*Spotlight* and *Sherlock*) and Windows (*Google Desktop* and *Windows XP Search*). They found that regardless of platform and search engine, users preferred navigation to search, and found no evidence that more searching took place in the "advanced" search engines. They found users were only willing to search when they forgot where a file was located, but were sure it was there somewhere. In other words, search is a "last resort". They wrote,

> "...in light of our findings, there is still no evidence that what prevents users from using search as a preferred retrieval strategy is the 'primitive' nature of current search engines; improving them did not change users clear preference for navigation in information retrieval." (p. 21)

Taken together, the results of these studies indicate that there is still a future for researchers studying organizing, labeling, and browsing. Recognition is a very powerful force in finding behavior precisely because it saves so much effort over recall, and it makes sense to expect that systems will continue to be designed to support browsing as well as search.

2.7 Summary

Group information repositories contain information that is shared by the users of the repository; however, this type of "sharing" is different from how files are shared using other types of systems. For example, in a group information repository there is no explicit sender and recipient; instead, a producer *packages* the information for future reuse by an unspecified information consumer. Packaging is a social process that involves imagining the context of the future reuse and "wrapping" the file in this additional information so that it can be found by the next person who needs it.

Packaging takes place in group information repositories in the form of organizing and labeling files. We know from the categorization literature that categorizing is a fundamental human process, and that if we are unaware of the similarities between items in a hierarchy we will automatically infer that they have something in common and try to figure out what that something is. In this way, users browsing a repository might actually learn about how other users think about and perceive the information in the repository.

Several studies have been published in the HCI literature about how people organize their personal information; there are fewer studies of organization in group information repositories. Regarding personal information organization, users find it to be a difficult task, and struggle to come up with categories that retain their relevance and usefulness over time as circumstances change. People rarely go back and reorganize information in personal or group information repositories, which means that early decisions when organizing information have a lasting influence on future use of the repository.

Labeling of information is also packaging for future reuse; however, a major obstacle to effective labeling for others exists. People just don't agree very often on the exact same words to label the same thing—in fact, they agree less than 20% of the time, and it requires considerable effort and incentives for them to overcome this tendency. Even when users of a group information repository agree to "naming conventions" when they start using a repository together, they can't stick to the agreement. Over time, they slip back into their idiosyncratic preferences.

Packaging is a social process, and as such theory from psychology and communications about how people achieve shared meaning and understanding might apply to packaging in group information repositories. Common ground is the mutual knowledge, beliefs and assumptions that people share about each other that accumulates via conversation and shared experience, and it helps people tailor their utterances so the listener can understand them; this is called audience design in the literature. Over time, this common ground is incorporated into a "model of others' knowledge"—a mental model of what one person can expect another person to know. We draw on these mental models when engaging in audience design.

Common ground is normally studied in the lab using "referential communication tasks", which are essentially tasks where two people must coordinate to accomplish some kind of task together, often involving creating or building something or solving a puzzle, without being in the same room. These tasks also have a labeling component, as participants need to be able to identify puzzle pieces (etc.) so that they can be sure they are both talking about the same thing. Because these studies take place in the lab, researchers can monitor and in some cases measure the common ground as it accumulates. The task of packaging for the future reuse of others is a kind of joint task involving coordination and naming task, although it is asynchronous to the extreme; I based the design of my thesis experiment on studies using the referential communication task paradigm.

Finally, recognition memory is extremely important for finding information in group information repositories. Users browsing a repository look for "cues" that signal to them that the information they are looking for might be in a particular folder, similar to information

foraging on websites. Users' preference for browsing relative to search is still very strong, despite advances in desktop search engines.

Chapter 3

Interview Findings and Log Data Analysis[1]

3.1 Introduction

In this chapter, I present findings from an interview study with users of a group information repository system. The purpose of the interview study was to describe information management tasks and behaviors resulting from social influences and constraints that are unique to group information repositories. Findings from the interview study indicate that social factors affect the information structure of the repository, and how it grows and evolves over time. Users restrict their activities to files they "own," are reluctant to delete files that might be useful to others, dislike the clutter that results, and can become demotivated if no one views files they uploaded.

3.2 Interview Method and Participants

CTools is a group information repository and course management system used at the University of Michigan, and is an instance of the Sakai open-source application platform in use at over 150 institutions worldwide. Users can create *project sites* which support storing and sharing files online. CTools project sites are increasingly popular: the number of active project sites grew 213% from December 2005 (3170 sites) to December 2007 (9932 sites). The interface for storing and sharing files is similar to that of many other web-based systems for group file storage: it follows the file-and-folder desktop metaphor, meaning that files and folders can exist in one and only one location in the hierarchy. There is no version control, access control exists at the site level, and there is no search functionality. Users can specify that a notification be sent when adding a file, and set preferences for receiving notifications when another user adds a file.

[1]The interview findings in this chapter were published as Rader (2009)

Table 3.1 Descriptives about the CTools project sites used in this study. Data reflect site characteristics as of the date interview sessions began.

Site	Type	Users	Site Age	# Files	Accesses/ Week
Site 1	extracurricular	26	2.5 yrs.	115	20.40
Site 2	staff/admin	89	1 yr.	452	76.08
Site 3	research group	11	4 yrs.	580	26.96
Site 4	research group	11	1 yr.	120	24.88
Site 5	research group	13	2.5 yrs.	894	22.73
Site 6	extracurricular	18	1.5 yrs.	407	19.10

Sixteen users of six different CTools project sites (4 men, 12 women) participated in semi-structured interviews[2]. The sites were selected for variety in the type of work each group conducted, and for activity level: each site had at least three members active on an approximately weekly basis. Respondents ranged from undergraduates to graduate students in several departments across campus to University staff members from different units who had been part of the organization for decades; some used CTools several times a week, while others did so rarely. At least two members of each site were interviewed, in their normal work environments wherever possible and in front of a computer so they could access the CTools site of their group as needed. Respondents answered questions about the work being conducted by the group, about their use of the CTools site, and about interactions with other group members. They also walked through the CTools site with the interviewer, describing the information available on the site and how it was organized.

The interviews were recorded, transcribed and coded using both inductive and deductive approaches (Miles & Huberman, 1994). Open, iterative coding took place in parallel with the data collection. Early coding stages consisted of labeling and categorizing participants' reported attitudes and interactions with respect to other group members and the information repository, and observations about the work practices and environments unique to each group. In the second stage of the analysis, I began with a set of types of activities users reported conducting with respect to their repositories, such as uploading, searching for, and deleting files. Finally, high-level themes emerged as connections were drawn between codes from the earlier stages of the analysis. When coding was complete, I summarized the data according to each high-level code using a matrix in the style of (Miles & Huberman, 1994) which helped me to identify patterns across participants and sites. The findings below report themes observed across all sites, despite differences in type of group, activity level, age and size (see Table 3.1).

[2]All names are pseudonyms. Some names of files and other details have been omitted to protect respondents' identities.

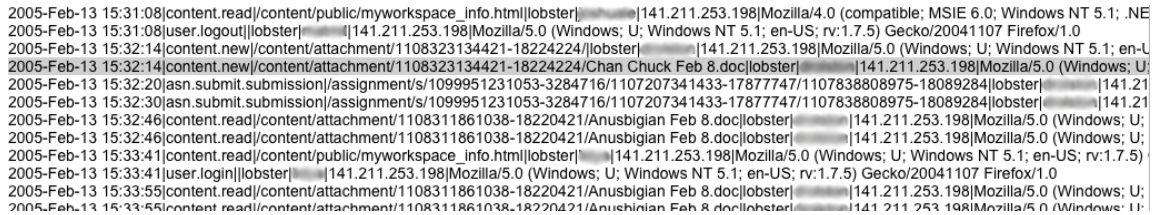


Figure 3.1 Excerpt from a CTools log file

3.3 Log Data Collection and Analysis Method

In addition to the interviews, I conducted an analysis of CTools system log data. Any time a user logs in, or creates (uploads), views, modifies, or deletes a document, a record of that action is captured in an event log file; an excerpt from a log file is presented in Figure 3.1. The log files were parsed into a MySQL database, users' unique identifiers were permanently disguised using the MD5 cryptographic hash function, and the database entries were cleaned of parsing errors and duplicate entries.

The purpose of the log data analysis was to "corroborate" findings from the interview study by looking for evidence of similar behaviors in data collected automatically by the system. Essentially, this means looking in system log data for actions that can be construed as proxies, or "behavioral traces" related to the themes I uncovered in the interviews. Starting from the system activity logs for every project site active from 2005-2008, I restricted my sample of sites to those that were similar in activity pattern to the sites included in my interview study. For example, recruiting for the interviews was restricted to "active" project sites, meaning sites that were used on average, by multiple different users in a given week. This same restriction was applied in sampling the log data. In addition, I restricted my sample to sites with between 2 and 100 members, and those active on more than 10% of days between the site creation date and the last activity recorded in the system logs for that site. From a total of 21,723 sites created in 2005-2008, this left me with a sample of 5,551 sites.

The system is only able to record evidence for actions users conducted within CTools. So, for example, if a user were to email a file to another site member using an email application rather than sharing it via the CTools site, that action would be an instance of file sharing that is not captured in the CTools log data. Also, users' attitudes and perceptions are impossible to detect in the log data, unless some action took place in the system as a direct result of those attitudes or perceptions. Even then, it would be necessary to make a convincing argument for why particular recorded actions could be considered a proxy for some attitude or perception. For example, when a user uploads a file to a CTools site, the system automatically records the name of the file as it appeared on the user's local machine.

In the upload interface, users are given the opportunity to create a new name for the file—a name that will be viewable by the other users of the site. The system does NOT record this user-viewable name as part of the log data. So any analyses I might imagine using filenames would be suspect, since the names available to me to analyze cannot be seen by the end users of the site.

3.4 The System: CTools

CTools is an instance of a group information repository system; it supports sharing documents and other files in a central online repository that project members can access. Users can create *project sites*, with which members can, "make announcements and share resources, such as documents or links to other resources on the web"[3].

CTools is an implementation of the Sakai open-source application platform currently in use by over 150 institutions worldwide, according to a presentation available on the front page of the Sakai website[4]. Figure 3.2 depicts the location of participating institutions[5]. Among the institutions using Sakai are Stanford University, Northwestern University, Indiana University, Georgia Tech and Yale University.

The *Resources* tool is similar in important ways to other systems used for sharing files. Users make explicit choices to store files in the repository, and must label them and add them to an existing file-and-folder hierarchy. Files can exist in one and only one location in the hierarchy. Unlike other systems, there is no search feature for CTools Resources. Like other systems, CTools has an underlying database which stores references to the actual location in memory of the items; however, users do not explicitly query the database. Instead, to access information in the repository, they must navigate the hierarchy structure and download the item before they can look at it. Users who have administrative access to a CTools site can restrict access by controlling the membership list for the site. Other capabilities of CTools include the ability to explicitly trigger a notification to be sent when a particular file has been added to the site; however, it is not possible to "watch" particular items or folders and be notified when they are updated. A screen capture of the Resources tool on a CTools project site is presented in Figure 3.3.

According to the UM CTools website[6], "Resources is the most widely used tool in

[3]https://collab.sakaiproject.org/portal/site/!gateway/page/!gateway-200

[4]http://bugs.sakaiproject.org/confluence/download/attachments/43394/newportbeach_20071204.ppt.zip?version=1, accessed on June 23, 2008.

[5]Screen capture taken July 12, 2009 at http://sakaiproject.org/portal/site/sakai-community/page/d89dabbf-a033-412f-80c4-a38931056b26

[6]https://ctools.umich.edu/portal/site/!gateway/page/1091327577225-1420520

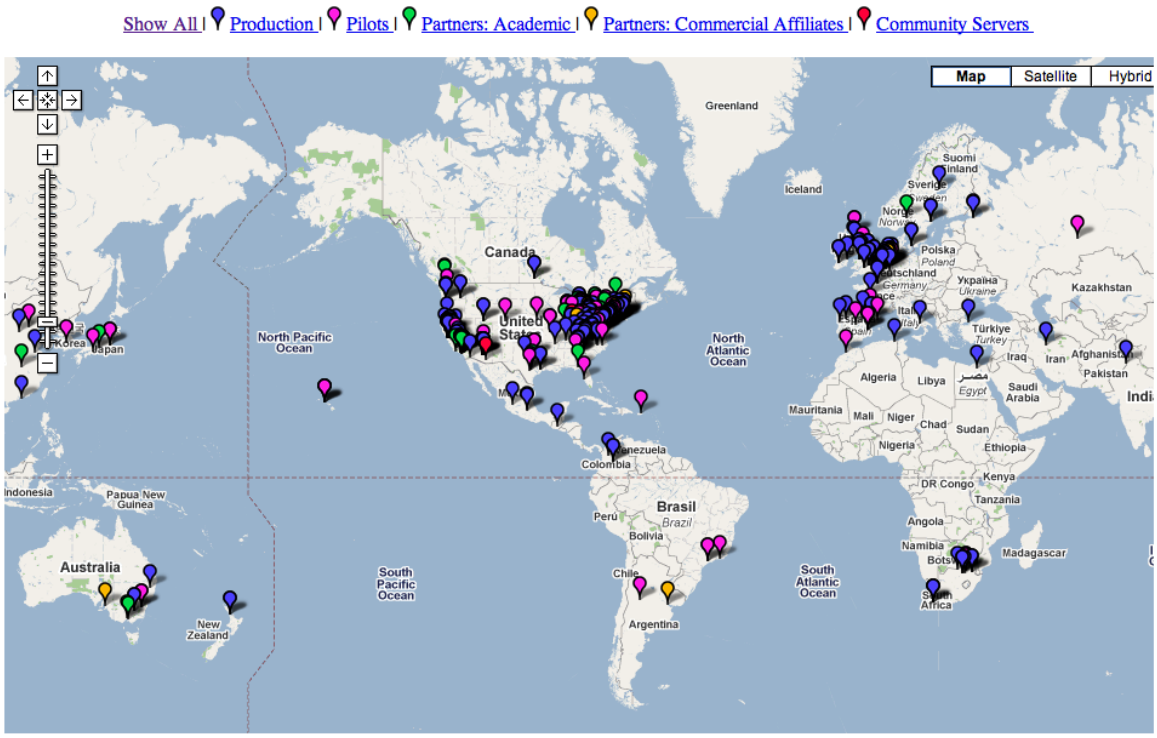


Figure 3.2 Map of the world depicting locations of institutions using Sakai.

Title		Access	Created By	Modified	Size
Resources	Add Actions				
2008 [redacted] days	Add Actions	Entire site	[redacted]	Mar 11, 2008 1:53 pm	2 items
Archived	Add Actions	Entire site	[redacted]	Feb 18, 2007 10:05 pm	4 items
2005 [redacted] - photos	Add Actions	Entire site	[redacted]	Feb 19, 2007 1:29 am	10 items
2004 Poster Session Photo	Actions	Entire site	[redacted]	Feb 18, 2007 10:09 pm	169.9 KB
IHMC CMap tool	Actions	Entire site	[redacted]	Feb 19, 2007 12:20 am	20 bytes
IM Handles	Actions	Entire site	[redacted]	Feb 18, 2007 11:39 pm	1.2 KB
Bios for [redacted]	Add Actions	Entire site	[redacted]	Aug 8, 2008 4:29 pm	1 item
Committee Assignments	Add Actions	Entire site	[redacted]	Aug 30, 2008 10:49 am	1 item
Consent Form Examples	Add Actions	Entire site	[redacted]	Jun 2, 2008 5:08 pm	9 items

Figure 3.3 Screen capture from the Resources tool on a CTools Project Site

classes and collaborations. In Resources, you can make many kinds of material available online. There are three main types: documents (word processing documents, spreadsheets, slide presentations, plain text, etc.); links to other websites; and documents that are created and displayed right on the CTools page."

3.5 Findings

Below, I present four major themes that were present across all of the sites from which respondents were recruited for this study. Wherever possible, I have supplemented the qualitative analysis with evidence from the log data.

3.5.1 *MY* Stuff vs. *YOUR* Stuff

When multiple information producers add files to a repository, complications can arise. As previous research has shown, it is very difficult to both initially agree upon and then subsequently adhere to conventions for how files should be labeled and organized (Berlin et al., 1993). This phenomenon was evident across all the sites in the interview study. For example, Nancy (Site 3) said,

> Probably the biggest problem we have with CTools is that people tend to organize information different ways, you know like you have a picture in your mind of how you think it should be organized, and that's not exactly how someone else's brain works to organize things.

Steve (Site 5) talked about trying to organize the site so that it would make sense to other people. He felt like he could only speculate on where other group members wanted things to be stored:

> I set up the directory structure and that kind of stuff. I'm always tweaking it. And I'm always looking for input, but I never get any input on stuff like, 'how would you like the stuff to be arranged'.

This tension leads to conflicts between "my stuff" and "your stuff" in a repository. Ideally, the contents of the repository are "our stuff", but in practice perceptions of ownership have very real consequences for repository structure.

Nancy and Zoe (Site 3) separately described the same incident that highlights consequences of this tension. Nancy was unhappy that members of her group all had different ideas about where things should be stored, but did not feel right about taking a hard-line approach to fixing the problem: "Unfortunately, I think it's important that people do things the same way but other people don't agree with me and I'm not an enforcer." However, when Zoe added a file to the site that Nancy felt should live in a different place, Nancy took action (described by Zoe): "So actually, I got an e-mail from Nancy telling me to move it to [a different folder]." When Zoe didn't move it, Nancy created a copy and put it in the new location herself without telling Zoe she had done so. During the interview, Zoe discovered her file in a different place than she had originally put it and said,

> It looks like she moved it to another folder but it's still here too... I'm the original one who posted it though, and it got moved around. But there's only two of us right now that are using it. And she has it in the place she wants it. And I have it in the place I put it.

Nancy took action to satisfy her own needs while preserving Zoe's original location and ownership of the file; however, this action also increased the clutter on the site. In another instance, Nancy described the history behind how two nearly identical folders on her group's site came to exist:

> And then [the folder called] 'Slides', this is where [the PI] sticks a bunch of powerpoint for talks that he's given on the whole concept of our lab... the 'Conferences' folder contains. Well. Yeah. Actually in a practical sense they should probably be the same. But the Conferences folder, contains stuff that's usually done by students, and the other Slides folder contains stuff that's done by [the PI].

Combining these folders might seem an obvious next step, and one that could even take place during the interview, but Nancy did not do it. The 'Slides' folder had been created by the PI, not by Nancy, and she was reluctant to make any changes to it. As a result, anyone needing presentations made by members of the group would have to look for the information in two places with different names.

Finding evidence for territoriality in the CTools event logs is challenging; a user might feel ownership toward or protective of a file created by someone else, for example, so simple signals like the identity of a file's creator cannot necessarily be interpreted as signaling territory. Instead, I chose to look for indications of parts of a given site in which the actions recorded by the system were dominated by one person. For example, I can count the number of "read" events within a particular folder, associated with each user of the site[7]. The higher the percentage of events in that folder associated with a single user, the more the activity in that folder is dominated by one person. Averaged across all folders and sites, 46% of events in a given folder were done by one person. In addition, 56% of the "read" events on a given file were done by the person who created the file. These percentages are broken out by the number of members per site in Table 3.2.

3.5.2 First Do No Harm

Information producers typically do not like to undertake the task of pruning the contents of a group information repository (Markus, 2001). Users feel they should hang onto information

[7] I did not count "new" or "modify" events in this analysis. "Read" events outnumber "new" events, on average per site, by 3.14 to 1, and "modify" events by 44.26 to 1

Table 3.2 Site "territoriality" statistics split into categories by the number of members per site (group size).

Group Size	Number of Sites	% Reads by a single user per folder	% File reads by the creator
Two	153	79.99	84.70
Three	495	60.25	66.42
Four	1183	49.88	58.49
Five	912	47.26	55.24
Six to Ten	1499	45.61	51.31
Eleven to Twenty	615	39.67	43.05
More than Twenty	694	27.07	30.56

they are not sure they need, just in case the need might arise later. Pruning decisions are especially difficult when repository contents are perceived as either yours or mine, but not "ours"; if users are bad at estimating their own future use, they are especially uncomfortable making assumptions about others' future information needs. The consequence of this uncertainty for respondents was they never deleted any files—even ones they had created! Josh (Site 6) said the only files he might feel comfortable deleting from his group's CTools site were outdated files he himself had created:

> *Josh*: I try not to [delete files]... so if another person goes looking for it, they can find it. The only times I delete, is when I post something more updated. And I'm sure that no one else will need the old version.
>
> *Interviewer*: How do you become sure that no one wants the old version?
>
> *Josh*: I only delete things I've posted.

However, Josh could not recollect a single specific instance when he had deleted a file. The only time Susan (Site 4) remembered deleting anything was, "...when we created the snack list, for last semester, and so I just created a new one for this semester and deleted the old one." When asked about what kinds of files she might delete, Linda (Site 2) said:

> Well, I can't think of anything. The only thing I can think of is if it's something really offensive, or doesn't pertain to [the project]. Then I would probably delete it. But I can't imagine anybody doing that. It hasn't really crossed my mind [to delete something].

Nancy (Site 3) recognized the value in periodically pruning the CTools site: "It seems like once a year somebody should sit down, like me or somebody else in the lab and kind of restructure all the new folders that pop up in here." Even so, when asked if such a

cleanup had ever happened, Nancy's response was "No [giggle]." While exploring the site during the interview, she came across at least three different folders that had been created years ago by members who left the group. The folders were subsequently abandoned, new group members had no way of knowing why they existed, and yet they remained on the site. Despite the potential for positive outcomes from pruning, respondents kept outdated files and folders around just in case they or someone else might need them. They were unwilling to make a decision that might directly prevent another member from accessing the information, especially if they were not the original creator of the file. Even these decisions NOT to remove something have consequences for the structure and evolution of the site.

The event logs provide evidence supporting Josh's claim. Of the 242,433 files both created and deleted from 2005-2008, 61% were deleted by the same person who created them, and 39% were deleted by someone other than the creator.

3.5.3 Consequences of Clutter

While there is no direct monetary cost or quota for sites that grow continuously and never shrink, users of cluttered and poorly organized repositories incur hidden costs in terms of lost time and effort. They may even be unaware of potentially important information: "It's not really people not knowing where to look, or not being able to find it once they're on there, but just not even knowing it's on there" (Zoe, Site 3).

David (Site 6) talked about how he was only familiar with the part of the CTools site where he kept files related to his part of the project: "So most of what I need is in the operations folder. To be honest I don't think I've ever really even looked in the marketing folder, because it's just not stuff that I need." Jennifer (Site 5) described the same phenomenon:

> See, you keep going and going and going and it's more folders and some more folders and some more folders, so if you don't really know, you're not really familiar with... let's see. 1,2,3,4,5 [levels]... yeah. So you really have to know your way around here or it can be quite intimidating. But like I know that my stuff is in here... wait a minute. Where is my stuff? Here it is! Like, I've been working on factor analysis, these are all my files. And I know that my reliability stuff is in here. So, I just know my spot, basically.

Jennifer did not explore other parts of the repository; she kept all the "stuff" she thought she needed in one place. However, because she did not visit other parts of the site regularly, she did not know what other information was available. During her interview, she came across a file she had not known was there: "And this one here... I totally need to read this, actually. This is really important, I should read this one. Because I'm doing work on that right now. It's my thing." Zoe (Site 3) had a similar experience during the interview:

> I don't really know what this is. It's probably just reference databases Nancy had on her computer and uploaded. I mean it would be beneficial for somebody to use... it looks like it hasn't been modified for a couple years. But some of it is interesting, and probably would have been useful for me when I was doing my quals.

As Jennifer and Zoe discovered, multiple people using a group information repository do not always know what other members might be adding, since they tend to restrict themselves voluntarily to the areas they are familiar with. In David's case, he voluntarily self-restricts because he is satisficing, only interacting with the parts of the site for which he has immediate need. The system provides access control only at the level of entire sites; one cannot programmatically hide individual files or sections of the hierarchy from specific members. And yet the self-restriction essentially creates the same kind of consequences.

Using average hierarchy depth as a proxy for "clutter", I was able to calculate Spearman correlation coefficients between measures of activity and the average depth for each of the 5551 sites in my sample. Deeper folders in the hierarchy have fewer files on average ($r = -0.37$), are accessed less often ($r = -0.34$), and are accessed by fewer users ($r = -0.26$).

3.5.4 Unmet Social Expectations

Respondents who added files to the group's CTools site expected that other group members would be aware of their contribution, and that some would make use of it. For example, Nancy (Site 3) talked about putting some items online that were directly related to something she was working on, that she thought others would want to see: "I was looking at reference papers of a co-PI who works with our lab and I posted them to our site because I thought they were papers that people should probably read and be aware of." However, the only way for her to tell whether anyone had accessed them was to ask directly, in person: "At our lab meeting I mentioned that I had posted these as well. But nobody said they had looked at them." Susan (Site 4) had a similar experience, posting manuals and research papers to the site. And Zoe (Site 3) posted her procedures for operating research equipment so that others could take advantage of them. Jennifer (Site 5) explained her thought process regarding whether to add a particular document she and another group member had been working on together:

> Doesn't it sound like it should be here? But the problem is that there are only like a couple of us right now that are really working on it, that really have our hands on it, and so we have our own versions. But yeah, this is important for... people should have access to that.

However, on the flip side, Frances (Site 2) spoke negatively of information posted by another group member: "[information] that [the group member] found, that she thought would be useful [for the group]". The implication of Frances' comment was that only the original poster, whom she called a "walking brainstorm", would find the information to be useful. Not knowing whether anyone else ever looks at the information one has added leads to the assumption that nobody is interested, and that can be demotivating. Linda (Site 2) experienced this, and talked about it in the interview:

> I did stop doing it [posting meeting minutes to CTools] for like a month, in January. The January meeting minutes are missing from CTools. I didn't do it and nobody said anything. And I'm like, why do I keep doing this?

The act of sharing a document via the repository carries with it an expectation that the audience is out there, despite the lack of a targeted recipient and the asynchronous nature of the potential exchange. When this expectation is violated in a group information repository it can reinforce the tendency to simply stick to one's own self-designated area of the site.

I did not find a suitable proxy in the log data to confirm this finding in other sites; however, this is observation has been studied and documented in other social systems (Erickson & Kellogg, 2000).

3.6 Discussion

In this chapter, I have highlighted four ways in which a group information repository can be shaped by social context. Users expect that when they upload a file to a repository, others in the group will see it; when this does not happen, it can be demotivating for the information producer. There are no explicit controls for how information should be structured or organized, and yet a structure emerges; users add their own files in the areas with which they are most familiar, regardless of whether that works for other members of the group as well. The social nature of group information repositories means that any decision to delete information can potentially affect everyone in the group, so evaluation decisions are much more difficult—so difficult, in fact, that deleting rarely happens and when it does, users only delete their own files. And finally, sites become so complex and cluttered that information consumers are simply unaware of files on the site that might be valuable for them to see.

Each finding above points to a way system designers can shift focus from the individual and improve information flow between users, to better support the social aspects of the group information repository system:

1. **MY Stuff vs. YOUR Stuff**: Users' knowledge of repository contents is necessarily biased toward what they've added themselves. Feedback supporting awareness of others' use of the system (Voida et al., 2006) is essential for instilling a greater feeling that the repository contains "our" information.
2. **First Do No Harm**: It can be difficult for individual users to differentiate useful, relevant files from older, outdated ones. The interface might provide information about which files are frequently accessed and which are not, serving as a basis for decisions about what to keep and what to delete. Further research is necessary to determine whether this information can empower users to overcome their reluctance to delete files "belonging" to someone else.
3. **Consequences of Clutter**: The system should provide support for identifying cross-folder commonalities and linkages when information producers make decisions about where in the hierarchy to store new files. This information could help users recognize connections between "my" stuff and "your" stuff, much the way that the *Frequently Bought Together* and *Customers Who Bought This Item Also Bought* fields on Amazon.com bring items to the attention of online shoppers.
4. **Unmet Social Expectations**: On the information consumer end, the system should provide easy access to a summary of the information available, and recent changes to the site. This approach is used successfully on Facebook to keep Friends in the loop on each other's status and activities. Without a summary, the only way to become familiar with files added by other group members is brute-force browsing of the entire site.

Group information repositories are different from tools for personal information management, and should be analyzed not just from the perspective of the functionality they embody and the information they contain, but also the social aspects of the context in which they are situated. This context influences the choices users make about where to store files, in essence constraining the way users package files in subtle ways that can be problematic as the site accumulates content over time.

Chapter 4

Experiment: Method

4.1 Introduction

In this chapter, I present the design of the experiment, and some of the background and tradeoffs that went into the design. The goal of this research is to understand how the influence of *common ground* and *audience design* on labeling and organizing choices in group information systems affects finding behavior. To test this, I designed an experiment that allowed me to detect potential performance differences when participants completed search tasks in file-and-folder hierarchies created by others with whom they shared common ground (or not), and tailored for different audiences.

Common ground is the mutual knowledge, beliefs and assumptions that people share about each other (Clark & Brennan, 1991). A person tailors his utterances to his communication partner, or *audience*, based on his assumptions and beliefs about what the other person knows; this is called *audience design*. People use this information when engaging in audience design in typical face-to-face conversation (Nickerson, 1999; Niederhoffer & Pennebaker, 2002). This theoretical framework has been applied to computer mediated communication (Clark & Brennan, 1991); however, it is not clear whether these fundamental properties of language use in communication also play a role when someone is labeling and organizing files that will be shared with others. Common ground helps people understand each other in conversation; might the same be true when the "communication" is mediated by a group information system?

One kind of common ground is *community membership*, shared by people who have characteristics in common but have never directly interacted. For example, two people who have lived in the same city but never met can be said to share community membership common ground (Nickerson, 1999). This would become apparent if one happened to bump into the other on the street and ask for directions—they would very quickly assess their

partner's familiarity with the area and tailor their utterances accordingly (Isaacs & Clark, 1987). I focus on this type of common ground in the experiment, and the effect of audience design when participants organize and label files for two groups of people with whom they share a greater or lesser amount of common ground.

4.2 Research Questions and Hypotheses

I conducted an experiment with three categorical independent variables: *Producer*, *Imagined Audience*, and *Consumer*. The *Producer* labeled and organized files into a hierarchy; for the experiment I recruited Producers from two different intellectual communities, such that some participants would share community membership common ground with each other, and some would not. Producers were instructed to tailor hierarchies for a particular *Imagined Audience*; levels of Audience were No Audience (or "None"), Self, and someone from the Same or Different community as the Producer. Finally, the *Consumer* searched for files in the a hierarchy created by a Producer; Consumers in the experiment could be from the Same or Different intellectual community as the Producer.

The research questions addressed in this experiment are:

1. How do common ground and intended audience affect file and folder labeling and organizing (i.e., packaging)?

2. How does the influence of packaging on the information structure affect finding behavior?

Based on the literature mentioned in the previous section, and relying heavily on the research design and results of Fussell and Krauss (1989), I can make the following predictions about search task performance under different combinations of the above independent variables.

Hypothesis 1: When the hierarchy *Producer*, the *Imagined Audience* for whom the hierarchy was tailored, and the *Consumer* are all from the same community, the Consumer will have the LEAST difficulty with finding.

Hypothesis 2: When the hierarchy *Producer* and the *Imagined Audience* for whom the hierarchy was tailored are from the same community, but the *Consumer* is not, the Consumer will have the MOST difficulty with finding.

Hypothesis 3: When the hierarchy *Producer* and *Consumer*, or the *Imagined Audience* and *Consumer* are from different communities, Consumers will have INTERMEDIATE difficulty with finding.

Hypothesis 4: When the *Imagined Audience* is *Self*, Consumers will have the LEAST difficulty if they are from the same community as the *Producer* and the MOST difficulty when they are from different communities.

4.3 Method

I conducted a two-part online experiment in which participants used a web-based application created specifically for the experiment, designed to closely resemble the familiar file and folder "desktop metaphor" user interface. I call this a online experiment because it was conducted entirely via the Internet, at the convenience of the participants. At no time did participants visit a lab or interact directly with the experimenter; all interactions were conducted by email, including incentive payments which were accomplished by sending Amazon.com gift certificates after participants had completed each phase of the experiment.

In the Organizing phase, participants created labels for a set of short text files, and organized them into a file-and-folder hierarchy. They were able to view the files online, edit file labels, create and delete folders, and drag and drop files into mutually exclusive folders (i.e., each file could exist in only one place). Figure 4.1 provides a screen capture of the organizing interface. In the Finding phase, participants later returned to the experiment application and completed a series of search tasks, in which they browsed hierarchies created by other participants to find specific files (Figure 4.2). The experiment server recorded detailed information about participants' interactions with the system that was later extracted and analyzed. At the end of each phase participants completed a post-questionnaire in which they rated the usability of the system, and answered questions related to the experiment task.

4.3.1 Participants and "Communities"

In this section I provide more detail about the participants in the experiment, to clarify ways in which the two community groups, Computer Science (CS) and Information Science (IS), were similar and different. The data in this section were collected via an online questionnaire that participants completed as part of the experiment procedure. Details about the questions participants answered, and complete descriptive statistics can be found in Appendix F.

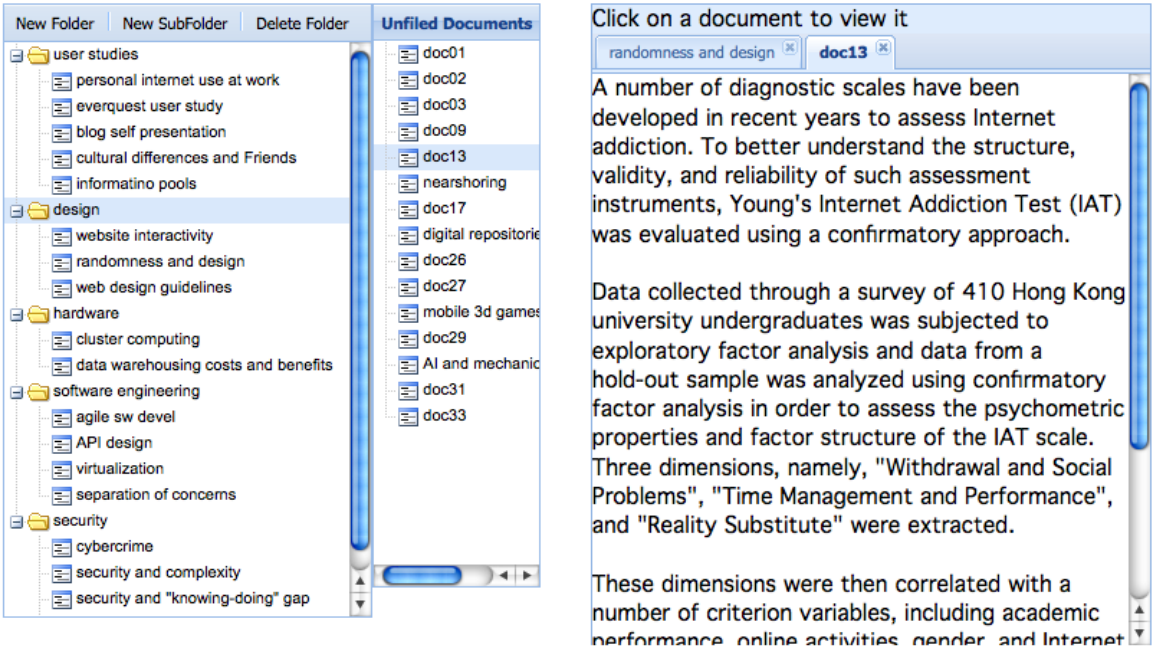


Figure 4.1 The organizing interface

Eighty-four graduate students participated in the experiment. Forty-one were students in the Computer Science department, and forty-three came from an Information Science department, both at the University of Michigan. In this experiment, "community membership" common ground is a subject variable; it is impossible for me to randomly assign graduate students to be members of either the CS or the IS departments. Subjects were, however, randomly assigned within community to each of the *Imagined Audience* conditions.

There are several key differences between the CS and IS graduate students as reflected in the questionnaire response data. For example, the gender of the participants was quite unbalanced across the two conditions. Three of the CS students were female; six of the IS students were male. In addition, 17 CS students reported that English was not their first language, while only 5 of the IS students reported the same. CS students also reported being exposed previously to slightly more of the topics in the files they organized (M = 2.24, SD = 0.80) than IS students (M = 2.81, SD = 0.79), where a score of 1 on a 5-point Likert scale meant they had been exposed to All of the topics, and 5 represented having been exposed to None of them. This difference was statistically significant (Kruskal-Wallis W=9.1916 (1, N = 84), p = 0.002). Finally, more IS students reported engaging in online "social media"-related activities, such as keeping a blog (72% vs. 54%), using Google Docs (95% vs. 73%), and using a shared calendar (98% vs. 78%), than CS students[1].

However, these differences do not represent a confound in the experiment; on the

[1]The results of questions about similar activities are reported on page 135, in Appendix F.

Table 4.1 Answers to question about what "might make sense to the target audience", by community membership condition

	M	*SD*	*N*
Same	2.63	1.16	20
Different	2.77	1.19	21
Self	2.09	0.90	23
None	2.00	0.86	20

contrary, they are examples of the kinds of characteristics that make these two different communities and are evidence that the *community membership* subject variable had distinct, measurable levels. Of more interest when considering possible confounds are questionnaire items designed to assess qualities like verbal ability, Internet use, past collaboration experience, and interest in reading the experiment files.

Regarding verbal ability, each participant answered five verbal analogy questions taken from Graduate Record Examination (GRE) practice tests available online[2]. The specific questions can be found on page 136. There were no significant differences between CS and IS students in the number of questions answered correctly (CS M=2.73, SD=1.07; IS M=3.05, SD=1.09, $F(1,48) = 1.43$, $p = 0.24$).

All subjects reported using the Internet between "Hourly or more often" and "Several times a day", and between "Some" and "Most" of the projects they were working at the time they participated in the experiment involved collaboration with others (CS M=2.63, SD=0.97; IS M=2.47, SD=0.83). In addition, When asked how interesting to read participants found the files they organized in the experiment, both groups on average gave neutral answers on a 5-point Likert scale that ranged from (1) "Strongly Disagree" to (5) "Strongly Agree" (CS $M = 2.83$, $SD = 0.97$; IS $M = 3.19$; $SD = 1.01$). There were no statistically significant differences between the conditions on these questions as well.

Finally, the questionnaire included two questions directly related to the experiment task and the audience design manipulation. After completing the organizing task, participants were asked how much they thought about what might make sense to someone in their target audience, if he or she had to find something in the hierarchy they just created (those in the "None" condition answered this question as if the target audience was "Self"). Responses were on a 5-point Likert scale, from (1) "A Lot" to (5) "Not at All". There were no overall main effects for *Imagined Audience*; however, pairwise comparisons between Different and Self (Kruskal-Wallis $W = 4.02$, $N = 44$, $p = 0.04$) and Different and None (Kruskal-Wallis

[2]http://www.ets.org/gre/; these questions were piloted as part of an experiment I conducted prior to my time as a graduate student at Michigan

Table 4.2 Answers to question about familiarity of members of the *Imagined Audience* with the topics in the files that were organized, by community membership condition

	M	*SD*	*N*
Same	2.42	0.51	20
Different	2.09	0.81	21
Self	2.82	0.78	23
None	2.15	0.48	20

$W = 4.69$, $N = 41$, $p = 0.03$) were significant. The means are presented in Table 4.1. The overall pattern shows that participants in the Same and Different conditions reported thinking about the target audience LESS than participants in the Self and None conditions, with the Different condition showing the least amount of thought.

The last question asked participants to assess how many of the topics or concepts in the files they thought someone in their target audience had been exposed to before. I compared the answers to this question with participants' self-assessment of their own familiarity with the topics. Responses were on a 5-point Likert scale from (1) All to (5) None; means and standard deviations for all conditions are reported in Table 4.2. There were no statistically significant differences by community membership (CS $M = 2.39$, $SD = 0.70$; IS $M = 2.37$, $SD = 0.76$), meaning that CS and IS students made similar guesses about how familiar their target audience was with the topics in the files. However, the Different and Self conditions showed statistically significant differences (Kruskal-Wallis $W = 7.83$ (1, $N = 44$), $p = 0.005$), as did the None and Self conditions (Kruskal-Wallis $W = 8.75$ (1, $N = 43$), $p = 0.003$). Overall, participants in the "Self" condition rated themselves as being exposed to fewer of the topics (remember, 5 = no familiarity) than participants asked to speculate about others (Different and Same). This indicates an overestimation of the knowledge of others, a phenomenon that has been documented in psychology research (Keysar & Henly, 2002; Wittwer, Nückles, & Renkl, 2008).

In summary, the results from the questionnaire data indicate that these two communities were different in several ways. It is reasonable to expect, based on their self-assessment of familiarity with the topics in the files, that the participants conceptualizations about how those files might be organized would not be identical across communities. In addition, an interesting pattern arose in the *Imagined Audience* condition regarding participants thinking LESS about what might make sense to the audience when organizing for some else, than they did when organizing for themselves.

Table 4.3 Organizing conditions, and number of participants

Imagined Audience	***Producer*: Computer Science**	***Producer*: Information Science**	**(total N)**
Same	9	10	(20)
Different	11	11	(21)
Self	11	12	(23)
None	10	10	(20)
(total N)	(41)	(43)	84

4.3.2 Text File Selection

The files used in this study were article excerpts, selected from recent issues of online periodicals and trade journals in the summer of 2008. A sample consisting of approximately 50 article excerpts were selected by the experimenter such that they all pertained loosely to current topics in Information and/or Computer Science. Some were more related to IS curriculum, some to CS curriculum, and some potentially interesting to both communities. The excerpts were also chosen to minimize the use of specialized vocabulary wherever possible—the topics were intended to be high-level enough that participants would spend their time and effort in the experiment organizing the files, not attempting to grasp the concepts in each of the texts. Thirty-three files were randomly selected from the sample for use in the experiment. The complete text of the files used in the experiment can be found in Appendix E.

4.3.3 Labeling and Organizing Procedure

In the Organizing phase, participants first completed a practice session to become familiar with the mechanics of using the interface (see Figure 4.1). Then, they read instructions that set up the experiment scenario. All were told to imagine they were writing a literature review paper; one-quarter (the "Self" condition) were instructed to organize the files so they could find them later if they needed to refer back to them. Instructions for participants in the "None" condition contained no reference to a potential future audience. The remaining participants were instructed to imagine themselves collaborating with someone from their own department, or with someone from the opposite department. So, for example, Information Science students were either told to assume they were working with other Information Science students, or students from Computer Science. The instructions participants received were similar to the below (complete experiment instructions can be found in Appendix B):

> On the following screen, you will be presented with a list of files. Each one contains a short article summary or excerpt. You may or may not already be familiar with the

> topics and concepts in the files. Your task is to create a more descriptive label for each one, and organize them into folders.
>
> There are many different ways to go about completing this task. Some people prefer to read through all of the files and create labels, before organizing them into folders. Others label a few at a time and create folders as they go, renaming and rearranging folders as necessary. What process to follow is completely up to you.
>
> When thinking about what to name the files and what folders to put them in, imagine that you are working on writing a literature review paper for a group project with Information Science graduate students at Large Midwestern University, and other members of your group will need to find some of the files later.
>
> In fact, Information Science students will be invited to participate in Part 2 of this experiment, and they may actually be asked to find files in the hierarchy you will be creating in this part of the experiment. So, please focus on creating a hierarchy with an organizational structure that would make the most sense for Information Science students.

Each participant, or *Producer* constructed a single hierarchy, for one particular *Imagined Audience*. See Table 4.3 for a depiction of the Organizing conditions, and number of participants in each condition. Participants were told they would not receive the incentive payment if they did not make a good faith attempt to organize the files; three participants were disqualified and replaced when it became apparent that they had not taken the organizing task seriously, from the spurious labels and extremely short time interval to complete the task.

4.3.4 Finding Procedure

About 60 days after organizing the files (M = 60.02 days, SD = 9.96), 48 participants revisited the experiment system and searched for a sequence of the same files they had organized, each one in a different hierarchy created by another participant.

Twenty-four participants in the Finding phase were Computer Science (CS) graduate students, and the remainder were Information Science graduate students.

Each participant completed two practice search tasks in the Finding interface (see Figure 4.2), and then 24 experimental search tasks. The experiment system displayed a target file, with no title, and a hierarchy with all of the folders closed; participants browsed the hierarchy, opening and closing folders and viewing files until they found the target. At that point, each participant (or *Consumer*) dragged and dropped the found file into a box at the bottom of the screen, and the next search hierarchy and search target were automatically displayed. Participants were not told who had created the hierarchy, or for what imagined audience. Because participants were able to view the target file for the entire duration of

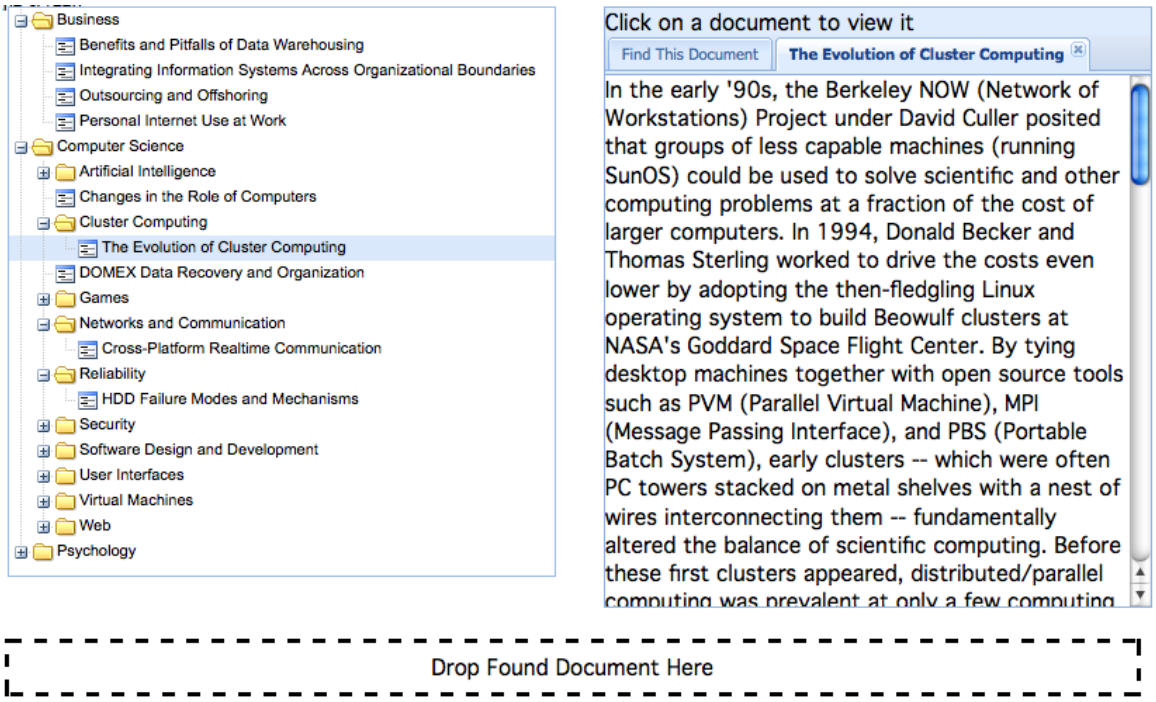


Figure 4.2 The search tasks interface

each search task, the finding success rate was nearly 100%. Similar to the Organizing phase, two participants who completed the Finding phase with an excessive number of incorrect searches in an unrealistically short time period (compared with to other participants) were disqualified and replaced with other participants.

4.3.5 Search Task Conditions

This study treats the hierarchies as communication artifacts, conveying information between the person who created it, the *Producer*, and the person searching within it, the *Consumer*. In addition, each *Producer* was instructed to tailor the artifact he or she created for a particular *Imagined Audience*, varied according to the Organizing phase instructions. We can conceptualize the relationships between pairs of these three real and imagined "interlocutors" in terms of the common ground they could potentially share. For example, if the *Producer* and the *Imagined Audience* are both IS graduate students, they are from the **same** community and therefore share some amount of common ground. Likewise, if the *Consumer* is a CS graduate student, they are from a **different** community and do not share much common ground. In another potential combination, a CS student is instructed to create a hierarchy for himself, and a different CS student searches within that hierarchy; here the *Producer* can be considered to share considerable common ground with the *Imagined Audience*, and the

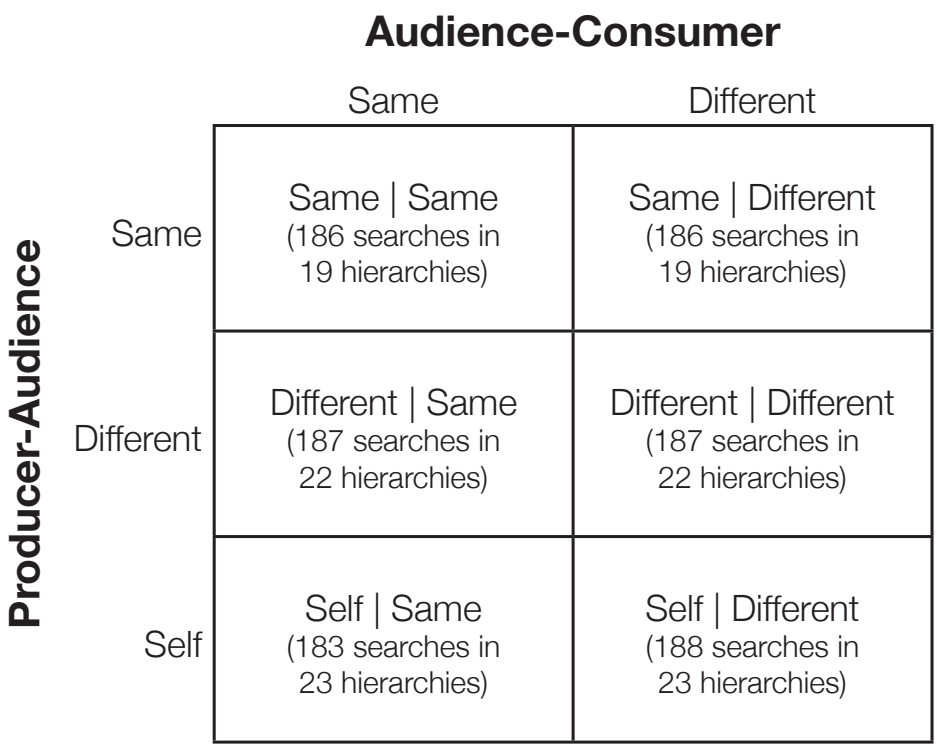


Figure 4.3 Conditions in the finding phase of the experiment

Consumer has common ground with both.

Figure 4.3 represents one way to combine pairs of interlocutors into two common ground dimensions by which the search tasks can be categorized: *Producer-Imagined Audience*, and *Imagined Audience-Consumer*. The figure has six cells corresponding to the search task categories. Participants in the Finding phase of the experiment searched for four target files in each of the six search task categories represented in Figure 4.3, for a total of 24. The files and hierarchy types were presented to half of the participants in one random order, and to the other half in a different random order, to check for potential order effects. Finally, there were 9-12 hierarchies that could potentially be searched-in for each search task category, corresponding to the number of hierarchies created in each condition of the Organizing phase (see Table 4.3). Finding phase participants completed search tasks in a subset of hierarchies, selected randomly without replacement, from within each category.

The "None" condition was included in the Organizing phase as a control to assess whether hierarchies created with an audience in mind differed in substantial ways from hierarchies created with no audience in mind. Preliminary qualitative comparisons of the hierarchies created in the "None" condition with other hierarchies showed no obvious effect of audience design. I subsequently chose not to include the "None" hierarchies in the search tasks, so as not to overburden participants with having to complete too many searches.

4.4 Rigor vs. Realism

The title of this section is inspired by a passage in Bae et al. (2006), who wrote about a study of what they called "document triage", or collecting, skimming, and organizing documents for personal use:

> "Studies of reading and organizing, even when they are apart, involve balancing concerns of rigor and realism. The known variability of individual reading and organizing practices suggests that we approach this balance creatively, giving participants a uniform task, corpus, and technology, yet allowing them to go about the task flexibly." (p. 8).

Tension always exists between external validity and generalizability, and this experiment was no exception. As I described in the introductory chapter, my high-level goal was to understand whether social factors and influences contribute to the difficulty people have in finding the information they need in shared information systems. For this thesis, I selected one type of system, the group information repository, which can be found in numerous different tools and systems for managing shared information and follows a user interface paradigm that is ubiquitous among knowledge workers. I borrowed theories from psychology and communications, and have attempted to apply them to help me understand what is going on in these systems. I wanted to produce results rooted in established theory that would help designers make better software. For that purpose, recruiting participants and having them do tasks that have some external validity is important.

As Bae et al. (2006) wrote, it is very difficult to achieve both realism and rigor in tasks that involve users working with documents. In order to exert some control over the experimental situation, it is necessary to hold the experiment materials constant across participants. However, this uniformity would most definitely never happen in the real world. In addition, early informal pilots of the experiment materials showed me that my estimation of how much people were willing to read, and the expertise level with the material I should aim at, were wildly inaccurate. In short, I initially expected way too much of people. Pilot participants quit in frustration before making it through a third of the files. So I made concessions for the sake of experimental rigor and feasibility, and tried to make the materials look as much like things my selected participants might be reading anyway as I could.

Similarly, in real group information repositories the information structure grows slowly, one file at a time, over years of use. In this experiment I chose to compress these "document triage" activities that people undertake in the real world (Bae et al., 2006) into a drastically shorter timescale. Some studies do not take external validity even this far; for example, in one of their labeling studies G. Furnas et al. (1983) had their participants generate labels

for a large number of recipes. Other researchers have generated a corpus of news articles, and screened out participants having prior experience with the topics, so they could hold background knowledge constant (D. M. Russell, Slaney, Qu, & Houston, 2006). These strategies were both contrary to my goals in this research—I am investigating the impact of that background knowledge, and participants assumptions about the background knowledge of others, as it relates to the topics in the files they organized and searched for.

Regarding the selection of computer science and information science graduate students as participants: before the experiment results were analyzed, it was not clear whether the communities in question would be *similar enough* to each other for interesting results to be produced by this design. But in hindsight, questions have been raised about whether the communities were too similar to have observed any community membership effects. My selection of the communities of participants was motivated by my belief that it is not outside the realm of possibility that computer scientists and information scientists might actually find themselves in the situation one day of collaborating on a project that uses a group information repository, and my need to select plausible files that would appeal to both groups based on my own knowledge and experiences. My primary goal was to identify communities that were different enough that it was possible they might not organize and label like the other group, and yet similar enough that they would have a chance of packaging appropriately if the intended audience was made more salient to them. This is a classic problem of floor vs. ceiling effects in designing experiments: too difficult, and all participants behave similarly because they cannot complete the task; too simple and everybody does so well there are no interesting differences. A small, informal pilot of the experiment did not produce consistent enough results to make this choice more clear, due to the large degree of variability across people in the way they organized and labeled the files. I believe it was both the similarities and differences between these two communities that made it possible to detect the interesting differences I will describe in the next chapter.

The selections of participants, documents, and experiment tasks allow certain kinds of generalizations, and prohibit others. For example, I cannot claim to have recruited a statistically random sample of people from the population of group information repository users, any more than a psychologist studying working memory can claim that college undergraduates participating for course credit are a statistically random sample representative of all humans (McNemar, 1946; Oakes, 1972). Likewise, because I did not select more than two communities from a spectrum of levels of community membership common ground, I cannot make claims about what might have happened in the experiment had I chosen theoretical physicists and romance language historians in addition to graduate students from computer science and information science. Nor did I measure the amount of common

ground between the two communities in the experiment to find out how similar they actually are; to my knowledge no such validated instrument for measuring common ground exists, and this is one of the reasons much common ground research is done in the lab with artificial tasks that researchers can observe and quantify the buildup of common ground (Schober & Brennan, 2003).

These limitations mean I cannot claim to have answered theoretical questions about the psychological processes of grounding and audience design. However, this was not my primary goal. I wanted to find out whether an information repository could serve as a kind of "communication medium" between people likely to use a system like this to organize material they find interesting and useful. I believe my emphasis on external validity means that my results can be interpreted in the context of similar tasks undertaken by similar kinds of people to those in the experiment. The wider applicability of research from any methodological tradition depends upon the arguments the researcher makes about the similarity between the situations and people they studied, and important aspects of the real-world phenomena they ultimately want to learn about. I will revisit this discussion of the impact these limitations of the experiment design have for interpreting and generalizing the results in Chapter 7 of this thesis.

Chapter 5

Experiment: Finding Phase Results

5.1 Introduction

In this chapter I present the results of the Finding phase of the experiment. I talk about this phase first, even though it took place chronologically after the Organizing phase, because the outcome of the Finding phase is the main result of interest in this research. The two-phase experiment was designed to allow measurement of the impact of packaging choices on finding behavior. The Organizing phase analysis and results will be discussed in the next chapter, as a way to understand what contributed to the pattern of results observed in the Finding phase.

5.2 Analysis Goals and Procedure

As described in the previous chapter, in the Organizing phase of the experiment 84 participants labeled and organized 33 files into file-and-folder hierarchies. Forty-eight participants returned later for the Finding phase and completed a total of 1138 search tasks using the aforementioned hierarchies, created under different *common ground* and *audience design* conditions[1]. The experiment server logged users' actions as they completed the organizing and search tasks, and these logs provided the data from which the measures for the experiment were constructed.

The dependent variable of interest in this experiment is the count of the total number of clicks ($\mathrm{total.clicks}$) required to find the target file in each of the search tasks. Smaller numbers of clicks mean better performance, i.e., participants were able to find the target file more easily, using fewer actions. Twenty-one of the 1138 search tasks yielded $\mathrm{total.clicks}$

[1]The hierarchies created in the "None" condition were not included in the Finding phase of the experiment, to reduce the number of searches required of each participant.

greater than 50 (M = 82.67); these values are remarkably extreme given that the mean file depth over all the hierarchies was 2.39 levels (SD=0.54), the mean number of folders was 9.55 (SD=4.05), and the mean folder size was 4.13 files (SD=2.21). These outliers were removed from the analysis, resulting 1117 total observations (20 to 24 per participant).

Analysis of Variance/Covariance is a common approach when analyzing data in which factors have been experimentally manipulated. However, I wanted to achieve four goals through this analysis, some of which are more easily accomplished using a generalized linear regression model:

1. Control for participant-, task- and hierarchy-level influences on the dependent variable, separately from the experimentally manipulated factors;
2. Conduct statistical hypothesis tests of the experimentally manipulated common ground and audience design factors;
3. Generate model predictions indicating the size of the differences between experiment conditions after controlling for other sources of variability (see #1);
4. And finally, compare these results against the theoretical predictions.

Participant-level influences, for example, are individual differences in working memory or familiarity with the topics represented in the files—things that might vary by participant and affect task performance, but were not explicitly measured. Task-level influences might include certain files being inherently easier to find than others, perhaps because of their uniqueness compared with the other files or their depth in the hierarchy. Hierarchy-level influences are things like the depth and breadth of the hierarchy, or the overall objective correspondence between the file and folder labels and the target documents.

I used the R statistical computing environment[2] to model the data using poisson regression. Poisson regression is more appropriate for count data like $\mathrm{total.clicks}$ because the assumption of normality is violated. The poisson regression model was estimated using maximum likelihood estimation, with a log link function and errors distributed according to a negative binomial distribution. The negative binomial distribution was necessary because these data are overdispersed; poisson regression yields standard errors that are too low, as well as a poor model fit. In using the negative binomial distribution, the model estimates an additional "dispersion parameter" and thereby compensates for the increased variability. Using the negative binomial distribution allowed me to perform more conservative statistical significance tests on the model estimates, reducing the probability of making a Type I error (Byers, Allore, Gill, & Peduzzi, 2003).

[2]http://www.r-project.org/, using glm.nb from the VR bundle

5.3 Regression Model

5.3.1 Predictors and Control Variables

The dependent variable in the model is total.clicks, the total number of clicks (consisting of all folder open, folder close, and file view events) to locate the target file. The regressors are:

- imagined.audience: the *Imagined Audience* for whom the hierarchy was created
- PA.Same: are the *Producer* and *Imagined Audience* from the same community? Yes or No
- AC.Same: are the *Imagined Audience* and *Consumer* from the same community? Yes or No
- imagined.audience * AC.Same: 2-way interaction
- PA.Same * AC.Same: 2-way interaction

The controls included in the model are:

- shortest.path: for each search task, the depth in the hierarchy of the target file, i.e., the absolute minimum number of clicks, or "shortest path" to find the target
- average.path.length: the average number of steps from any file in a hierarchy to any other file, used as an indication of the complexity of the hierarchy; for example, a hierarchy with files grouped into only two folders at the same level has a lower average.path.length than a hierarchy with 4 or 5 levels and fewer files per folder
- consumer.id: because each person experienced all types of search tasks, the model includes a fixed effects control for individual differences

An alternate way to account for the influences of variations in the shortest.path across search tasks might be to use the difference between total.clicks and shortest.path as the dependent variable. However, this strategy does not account for potential "side effects" due to the level at which a particular file resides in the hierarchy. For example, a file three levels deep in the hierarchy (shortest.path = 3) might require just 2 more clicks to find than a file that is one level deep in the hierarchy (shortest.path = 1). However, it is also plausible that to find a file buried three levels deep in a hierarchy, a participant might end up clicking around for a while in a first or second level folder before making their way down to the third level. In that case, a shortest.path of 3 signifies a task that might be of greater complexity than simple subtraction can account for. In other words, the "ideal" shortest.path might underestimate the difficulty as measured by total.clicks; the actual shortest path to a target

three levels deep might not be so direct. Incorporating $\mathrm{shortest.path}$ as a control in the model makes it possible to capture these complexities and estimate their actual influence on $\mathrm{total.clicks}$.

In within-subjects experiment designs like this one, there are two ways one can control for variation due to participant individual differences: fixed effects or random effects. The goal of both statistical techniques is to account for any nonrandom influence on the dependent variable that may be due to some latent, unmeasured construct. For example, some participants might naturally be better at searching than others. The experiment did not include a measure of "search skill", because it is not theoretically relevant to the research questions at hand; however, it is important to control for such individual differences so that the effects of the theoretically interesting predictors can be better identified.

With fixed effects, a dummy variable is included in the model for each participant, allowing for the precise estimation of an intercept for each person. A loose interpretation of this intercept is that it represents the baseline number of clicks to find the search target for each participant in the experiment. The inclusion of a coefficient in the model for each participant means the number of degrees of freedom can become very large very quickly, so the precise estimation gained is offset by a loss of statistical power (the more degrees of freedom, the more likely a Type II error).

With random effects, the regression equation includes just one term representing an estimate of the effects due to predictable individual differences plus random error; this allows for greater statistical power. In using random effects, the researcher assumes that the participants have been randomly selected from a much larger population, and that the effects due to participant-level individual differences are normally distributed (Gujarati, 2003, p. 650).

I chose to use fixed effects in the model, for several reasons. First, participants in this experiment were not randomly drawn from a larger population—as I discussed in Chapter 4 on page 53, this is not a statistically random sample. I do not claim the participants in this experiment are representative of the population of all group information repository users, therefore the first assumption for random effects does not apply.

Second, there is no evidence that the "individual differences" in question here are normally distributed; by definition, any variability lumped into "individual differences" is unmeasured and therefore it is difficult to verify this assumption (Verbeke & Lesaffre, 1996). Therefore, the second assumption for random effects is suspect. This is an important consideration, because bias can be introduced in the values predicted by the model by forcing the individual differences to fit a normal distribution when they otherwise would not (Verbeke & Lesaffre, 1996; Ghidey, Lesaffre, & Verbeke, 2008).

Third, in a "mixed model" that includes both fixed and random effects and a non-normal distribution assumption for the dependent variable (which is what this model would become if I were to use random effects for consumer.id), determining the appropriate degrees of freedom for the random effects is currently a hotly debated question in statistics[3]. This makes hypothesis tests of the coefficients impractical. In other words, I would be able to produce coefficients for the predictors and test statistics using a "mixed model", but I would not be able to find an appropriate distribution from which to calculate p-values (Bolker et al., 2009). This kind of statistical result was one of the stated goals of the analysis, so fixed effects are more appropriate here. And indeed, as later sections will show, the loss of power resulting from the use of fixed effects was not a problem for this analysis.

Finally, it is important to note that coefficients in a model estimated using fixed effects are not incorrect or biased (Gujarati, 2003). However, coefficients in a model estimated using random effects might be, especially if the normality assumption is violated. It is difficult to imagine what distribution "individual differences" might take, especially when these influences are undefined by design. But, for example, what if participants had different "persistence" thresholds for how long they would keep clicking in order to find the search target? There is some evidence that this was the case, given that the maximum number clicks required to find any search target was 183, before outliers greater than 50 were removed for the analysis. The (unmeasured) value of this threshold might not be normally distributed in a population. Because using fixed effects produces unbiased estimates, I chose to use that technique here.

Ultimately, this model is a tool used to understand what is going on in this particular dataset; any implications or broader applicability based on the results presented here does not hinge on whether fixed or random effects were used. Whether or not these results can be applied to situations beyond this experiment depends upon how believable this experiment is as an approximation of the real world. Anyone using these results should consider the context in which the data were collected (sampling frame, selection of experiment documents, task details, etc.), and how similar that context is to their area of interest, like they would with any other applied research findings. As in all experimental research, whether or not the results are generalizable is more a research design issue than a statistical issue.

[3] See the discussion at https://stat.ethz.ch/pipermail/r-sig-mixed-models/2009q1/001802.html

5.3.2 Model Specification

The model is constructed as follows:

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{imagined.audience},\\ &\text{PA.Same}, \text{AC.Same}, \text{imagined.audience} * \text{AC.Same},\\ &\text{PA.Same} * \text{AC.same}, \text{consumer.id})\end{aligned}$$

Deciding what regressors to use in a model like this is an iterative process that involves specifying the model with different combinations of predictor variables and comparing the model variations using likelihood ratio (LR) tests. LR tests compare deviance, which is a goodness-of-fit indicator, between different models (Agresti, 2007). I performed two LR tests to compare the goodness of fit between the final model, specified above, and three variations. First, I compared the model above with the "saturated model", a statistical construct that includes one regressor for every observation. As such it is able to perfectly predict the observed values. The LR test against the saturated model tests the null hypothesis that the saturated model and less-specified model are effectively the same. If this test results in a p-value above the threshold for significance, the null hypothesis is retained, and the less-specified model is an adequate fit for the data. For the LR test comparing the above model against the saturated model, the p-value was 0.32, indicating that the less-specified model is a reasonable fit. Similarly, a likelihood ratio test comparing the final model with a model including all possible two- and three-way interactions was not significant (p = 0.23).

Finally, there were two predictors that could impact the results of the experiment, but were not included in the model: the length of the time interval between the Organizing phase and the Finding phase, and which random order of hierarchies and search target files participants viewed. These two variables are highly collinear with the consumer.id regressor, and because consumer.id is a necessary part of the model it is unwise (and unnecessary) to include these other collinear regressors. The model estimates are accurate regardless; however, it is not possible for the model to estimate the amount of potential influence of either time interval or random task order on the dependent variable if they are not included as regressors.

Instead, I calculated Pearson correlations to assess whether these variables were related to the total number of clicks to find the search target. The correlation between time interval and total.clicks is 0.0006 (t = 0.021 (1115, N = 1117), p = 0.9832), and the correlation between task order and total.clicks is -0.009 (t = -0.3076 (1115, N = 1117), p = 0.7584). For the sake of comparison: the correlation between shortest.path, one of the control measures included in the model, and total.clicks is 0.21 (t = 6.9958 (1115, N = 1117), p $<$

Table 5.1 Negative Binomial Regression estimates, % Change, and Std. Error. Theta (dispersion paramater) = 2.728. consumer.id dummy variable coefficients are included in Appendix I.

	Regressors	*Estimates*	*% Change*	*Std. Error*[4]	
0.	(Intercept)	0.744	(2.10 clicks)	0.244	**
1.	shortest.path	0.189	20.862	0.048	***
2.	average.path.length	0.229	25.740	0.056	***
3.	imagined.audience (Info. Sci.)	−0.012	−1.217	0.130	
4.	imagined.audience (Self)	0.134	14.351	0.114	
5.	PA.Same (Yes)	−0.183	−16.749	0.089	*
6.	AC.Same (Yes)	−0.062	−6.009	0.150	
7.	imagined.audience (Info. Sci.) * AC.Same (Yes)	0.0956	10.035	0.226	
8.	imagined.audience (Self) * AC.Same (Yes)	−0.167	−15.384	0.191	
9.	PA.Same (Yes) * AC.Same (Yes)	0.037	3.771	0.125	

* p < .05; ** p < .01; *** p < .001

0.000). These correlations indicate that time interval and task order did not influence the outcome of the experiment.

5.4 Results

The regression results are detailed in Table 5.1. The Regressors are the explanatory variables and controls included in the model. For each regressor, there is an estimated coefficient. Remember that this model uses a log transform of the dependent variable; the model predicts the log of the total.clicks to find a search target rather than the actual count. The estimated coefficients are in the same units and must be transformed back before they can be easily interpreted.

Consider the estimate for shortest.path. If the shortest path to the target file were to increase by one unit, the difference in the log of the expected number of clicks is predicted to increase by 0.19 units, holding other regressors in the model constant. The estimate represents the log of the ratio of the expected number of clicks when shortest.path is 0 vs. when it is 1. This is difficult to conceptualize in terms of quantity of impact on a particular search task. Interpretation is made easier by transforming the estimate to represent a percentage change in total.clicks for every 1-click difference in the shortest path length.

[4]The studentized Breusch-Pagan test was significant (B = 91.39, df = 56, p = 0.0018), which indicates that heteroskedasticity is present and standard errors are likely underestimated. Table 5.1 reports White's robust standard errors (Gujarati, 2003) which are more conservative in the presence of heteroskedasticity; Wald tests on the estimates reflect the adjusted standard errors, reducing the probability of Type I error.

Calculating the percentage change is fairly simple: exponentiate the estimate, subtract 1, and multiply by 100. For shortest path, this yields 20.86%. So now we can say that for each 1-click increase in the shortest path to reach the target document, the total number of clicks to find the search target increases by 20.86%.

A Wald test was performed on each estimate to test the null hypothesis that true estimate of the coefficient is zero. The Wald tests allow me to test experimental hypotheses, similar in logic to the *F*-test in ANOVA. These significance tests are most interesting for the experimentally manipulated factors: imagined.audience, PA.Same, and AC.Same (and the interactions). Table 5.1 shows that the only significant estimate is PA.Same. The Intercept, and two controls in the model are also significant (shortest.path and average.path.length). The lack of significance of the other estimates does NOT mean those estimates are somehow biased or less reliable; it simply means that in the context of this particular set of regressors included in the model, we cannot statistically conclude that the actual coefficient is different from zero.

5.5 Model Interpretation

Transforming the model estimates into percent change is only the first step in interpretation. What does a 16.75% decrease in total.clicks for PA.Same, holding everything else in the model constant, really mean in practice for participants in the experiment? Estimates and percent change values for the regressors must be interpreted in the context of the rest of the model, and that means starting from the Intercept. Because most of the regressors in this model are categorical, the concept of these regressors having a value of zero does not really make sense. So instead of thinking about the Intercept as the value of the dependent variable when all other coefficients are zero, it is actually the total.clicks for a particular combination of the categorical variables selected to be the baseline by the model.

For this model, the base rate of clicks per search task is 2.10 when shortest.path and average.path.length are zero, and the categorical regressors take on the values of *Producer-Audience* (Different), and *Audience-Consumer* (Different). Figure 4.3 on page 52 depicts the possible combinations of the categorical regressors. One additional categorical dimension can be layered on top of these: *Imagined Audience* type. The Intercept takes on the value imagined.audience (CS); other levels of this categorical regressor that appear in Table 5.1 on page 62 are (IS) and (Self).

Table 5.2 on page 64 presents the results of the model, interpreted not as coefficient estimates but as differences from the Intercept, and compared with theoretical predictions

Table 5.2 Model Results compared with Theoretical Predictions. Model results are presented as % Change (from the Intercept) in total clicks to find the search target; "Best" means fewest clicks.

	Regression Model Results		$\Rightarrow$	*Theoretical Predictions*	*Producer & Consumer*	*Audience & Consumer*	*Producer & Audience*	*Imagined Audience*
1.	-22.39%	Best	$=$	Best	Same	Same	Same	Self
2.	-18.80%	Best	$=$	Best	Same	Same	Same	Comp. Sci.
3.	-11.74%	Best	$=$	Best	Same	Same	Same	Info. Sci.
4.	-1.22%	Worst	$\downarrow$	Intermediate	Same	Different	Different	Info. Sci.
5.	(Intercept)	Worst	$\downarrow$	Intermediate	Same	Different	Different	Comp. Sci.
6.	-6.01%	Intermediate	$=$	Intermediate	Different	Same	Different	Comp. Sci.
7.	$+2.16\%$	Worst	$\downarrow$	Intermediate	Different	Same	Different	Info. Sci.
8.	-4.80%	Intermediate	$\uparrow$	Worst	Different	Different	Same	Self
9.	-17.76%	Best	$\uparrow$	Worst	Different	Different	Same	Info. Sci.
10.	-16.75%	Best	$\uparrow$	Worst	Different	Different	Same	Comp. Sci.

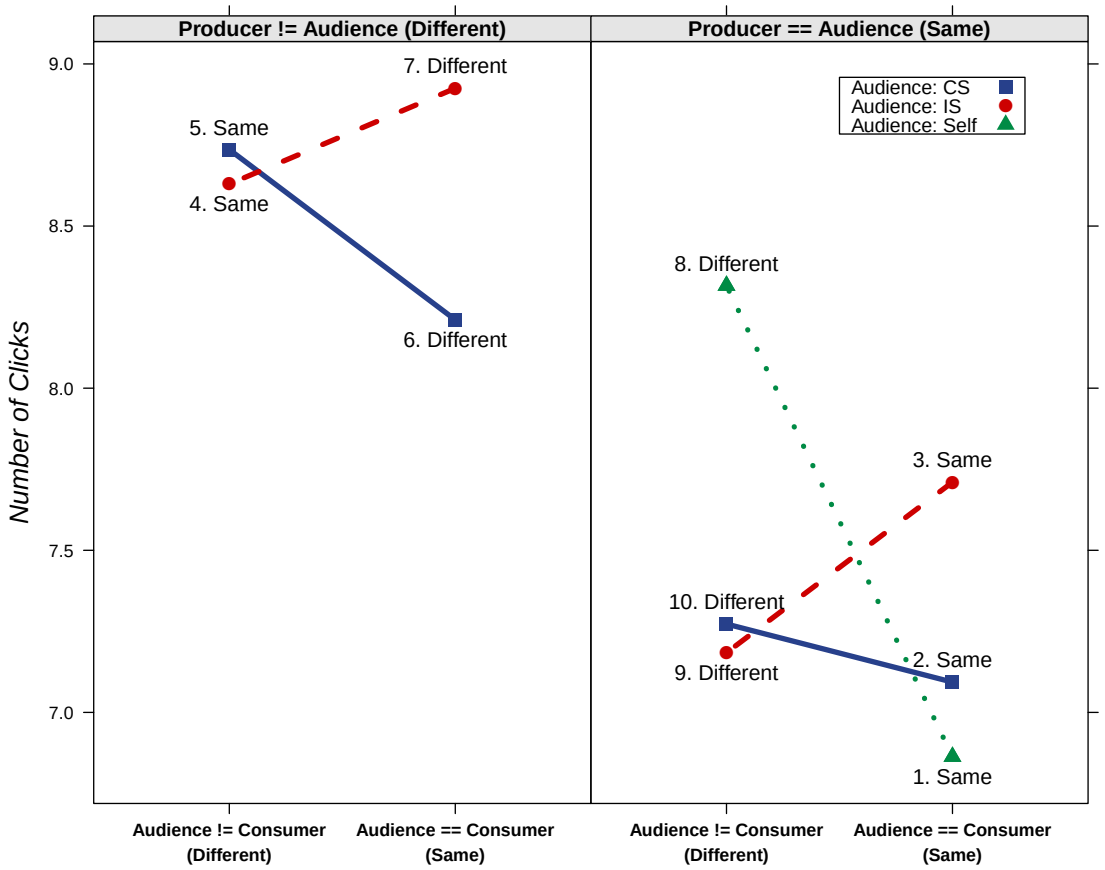


Figure 5.1 The regression model results, represented as fitted values, based on the median consumer.id estimate and mean shortest.path (2.48 clicks) and mean average.path.length (3.93 clicks). "Same" and "Different" point labels refer to *Producer* & *Consumer* community membership, and row numbers in Table 5.2. The lines on the graph illustrate the three *Imagined Audience* conditions: CS, IS, and Self. The left and right panels represent whether or not the *Producer* and his *Imagined Audience* are from the same community. Within each panel, the left and right depict whether the *Imagined Audience* and *Consumer* are from the same community. There are three things to notice about this graph. First, the difference between the left and right panels corresponds to the only significant model estimate for an experimentally manipulated factor (PA.Same). Regardless of the *Consumer* community, participants performed better (fewer clicks) when the *Producer* believed the *Imagined Audience* was similar to them. Second, the model predictions for the Self condition replicate the findings of Fussell and Krauss (Fussell & Krauss, 1989). Finally, the base rate of clicks to find the search target, for mean shortest.path and mean average.path.length and holding all other factors in the model constant, is 6.41 clicks. (Also note that the y-axis starts around 6.75, not zero.)

based on the literature. The rows in the table correspond to the ten points in the fitted values graph (Figure 5.1, page 65). The Intercept, row 5 in the table, is the baseline against which the percent change values are added or subtracted. The percent change numbers come from combining the appropriate model estimates of the effect of different levels of the categorical regressors. Consider, for example, the model prediction of $+2.16\%$ (row 7 in the table). This number is calculated by adding the estimates from Table 5.1 for $\mathrm{imagined.audience}$ (Info. Sci.) in row 3, $\mathrm{AC.Same}$ (Yes) in row 6, and $\mathrm{imagined.audience}$ (Info. Sci.) * $\mathrm{AC.Same}$ (Yes) in row 7, and then transforming them into a percent change. The rest of the percentages in that column can be calculated in a similar manner.

Looking at the fitted values in Figure 5.1 (page 65), it is clear that in the context of this experiment, the percentage differences between search task conditions translated into at most a 2-click difference between the highest and lowest points in the graph. This raises a question about statistical vs. practical significance; statistical magic aside, is this difference really large enough to be important? My argument is yes. A real-world group information system would likely be broader and deeper and contain more files than the hierarchies created in this experiment, increasing both the shortest path to the target file and the overall complexity of the hierarchy. This is reflected in the model estimates, translated to percentage change, for $\mathrm{shortest.path}$ (21% increase in $\mathrm{total.clicks}$) and $\mathrm{average.path.length}$ (26% increase in $\mathrm{total.clicks}$).

In general, Consumers performed best (fewest clicks to find the target file) when the *Producer* created a hierarchy for an *Imagined Audience* from the same community, regardless of the community the *Consumer* was from. Consumers had the the most difficulty when searching in hierarchies created by a *Producer* for an *Imagined Audience* that was not like them. This is an interesting and unexpected result; it means that who the Producers THOUGHT their audience was, turned out to be more important than who the Consumers ACTUALLY were. Said another way, Producers created hierarchies in which everyone could find stuff more easily, regardless of what community they were from, but only when they imagined that they were organizing for somebody like them. When *Producer* == *Imagined Audience*, all Consumers found the target in fewer clicks, regardless of whether they were like the Producers, or members of the target audience category (see Figure 5.1).

5.6 Hypotheses, Revisited

The model results can be used to evaluate the hypotheses outlined at the beginning of this paper:

Hypothesis 1: When the hierarchy *Producer*, the *Imagined Audience* for whom the hierarchy was tailored, and the *Consumer* are all from the same community, the Consumer will have the LEAST difficulty with finding. This prediction says that the best possible situation for a Consumer is to search in a hierarchy created by another Producer like them, tailored for someone from the same community. In this case, common ground is shared all around, and audience design is easy. This hypothesis was **Confirmed**. Rows 1-3 in Table 5.2 show that the fewest number of clicks were required in search tasks with these characteristics.

Hypothesis 2: When the hierarchy *Producer* and the *Imagined Audience* for whom the hierarchy was tailored are from the same community, but the *Consumer* is not, the Consumer will have the MOST difficulty with finding. The logic behind this is that both common ground and audience design work against the Consumer, who is from a different community. Surprisingly, this hypothesis was **Rejected.** Rows 8-10 in Table 5.2 show that where the literature predicted the worst performance, Consumers experienced some of their best performance. This is due to the *Producer-Imagined Audience* effect described above.

Hypothesis 3: When the hierarchy *Producer* and the *Consumer*, or the *Imagined Audience* and *Consumer* are from different communities, Consumers will have INTERMEDIATE difficulty with finding. Under these circumstances, it was expected that despite the *Producer* being instructed to tailor the hierarchy for someone from the opposite community, the common ground and audience design effects would be more important. This hypothesis was also unexpectedly **Rejected**. Rows 4-7 in Table 5.2 show that Consumers experienced the most difficulty under these search task conditions; the *Producer-Imagined Audience* effect is apparent here as well. All four rows say "Different" in the *Producer* & *Imagined Audience* column.

Hypothesis 4: When the *Imagined Audience* is *Self*, Consumers will have the LEAST difficulty if they are from the same community as the *Producer* and the MOST difficulty when they are from different communities. The prediction says that when a Producer customizes a hierarchy for herself, a Consumer from the same community uses 17% fewer clicks to find the target file than a Consumer from the opposite community (rows 1 and 8 in Table 5.2). This hypothesis was **Confirmed**, and is a replication of the Fussell and Krauss experiment (Fussell & Krauss, 1989).

5.7 Discussion

This research was conducted with three goals in mind. The first objective was to determine whether communication processes are at work in group information management tasks. Given the results of this study, it is clear that the answer is yes. This suggests thinking about labeling and organizing not just as storage and categorization, but as a communicative activity.

The second goal was to better understand influences of common ground and audience design on hierarchy creation and finding behavior, while replicating previous work. This goal was also accomplished. The main finding of this work is that all participants searched more efficiently in hierarchies created by Producers who organized for an audience they believed to be similar to themselves ("Same"), regardless of whether they were from the same community as the Producer, or were actually in the target audience. In contrast, when Producers tailored their hierarchies to dissimilar others ("Different"), Consumers required the most clicks to reach the search target. Participants performed best when they were from the same community as the Producer, were a member of the target audience, and the Producer organized for "Self". Replicating previous findings lends credibility to these results, and provides confidence that what happened in this experiment is indicative of larger patterns rather than local variations.

This raises the question, why might some of these results differ from predictions based on the literature? Consumers underperformed expectations when Producers tailored their hierarchies for different others, and did better than expected when Producers organized for similar others. There are two possible reasons for this. The audience design condition was more nuanced in this experiment than in Fussell and Krauss (Fussell & Krauss, 1989); rather than just "self" and "other", this experiment had two different flavors of "other" depending on the community membership of the participant. Providing participants with more information to use when doing audience design could easily have allowed for greater nuance in the results. Also, the instructions participants received referred to real communities, and the organizing and labeling task has greater external validity.

The third goal of this research was to gain insight into ways we might design systems that incorporate better support for social aspects of group information management. Given the *Producer-Imagined Audience* result, and in light of previous audience design research, I suspect that finding ways to incorporate support for the formation of more accurate mental models of other users could help. For example, Wittwer et al. (Wittwer et al., 2008) found that the level of detail in experts' models of laypersons' knowledge was important for successful communication; mental models that were either oversimplified or too complex

proved to be less effective. The same might be happening in this experiment, and would help explain the pattern of results. However, because I did not experimentally manipulate participants' mental models of the audience, making a stronger claim about this is left for future work. Also, this new audience design hypothesis does not fully explain why ALL participants who searched in hierarchies by Producers that organized for similar others did better; this means, for example, that both CS and IS people had an easier time searching in hierarchies created by a CS person who imagined their audience was similar to them. Analysis of data from the Organizing Phase, presented in the next chapter, will shed more light on why this pattern of results may have occurred.

Chapter 6

Experiment: Organizing Phase Results

6.1 Introduction

The goal of the experiment and analysis described in Chapter 5 was to discover how communication processes might play a role in group information management tasks. Recall the details of the experiment design: an information *Producer* labeled and organized a set of 33 files into file-and-folder hierarchies[1]. Each *Producer* was a graduate student in Computer Science or Information Science; this difference represents the two levels of the *common ground* independent variable in the experiment. Each *Producer* was instructed to imagine they were working on a literature review paper, and to organize and label the files such that someone from a particular *Imagined Audience* group would be able to find specific files later—this was the *audience design* independent variable. The conditions of the experiment can be found in Figure 6.1; instructions for participants in the "None" condition did not include an *Imagined Audience*.

The Finding phase results indicated that there was an effect of *audience design*: when a *Producer* believed s/he was creating a hierarchy for someone similar to themself, all Consumers performed better, regardless of the *Consumer* community. However, it is impossible to judge from those results alone what specific effects audience design had on the hierarchies that caused the differences. To better understand the audience design effect, I conducted a series of quantitative analyses of the hierarchies created in the Organizing phase. In this chapter, I make the argument that these differences were were due to variation in label choices when participants were creating the hierarchies, depending on who the *Imagined Audience* was. The statistics I report in this chapter primarily focus on the *audience design* independent variable; any results for the *Producer community* variable will be presented where they are relevant.

[1] see Appendix E for the text of the files

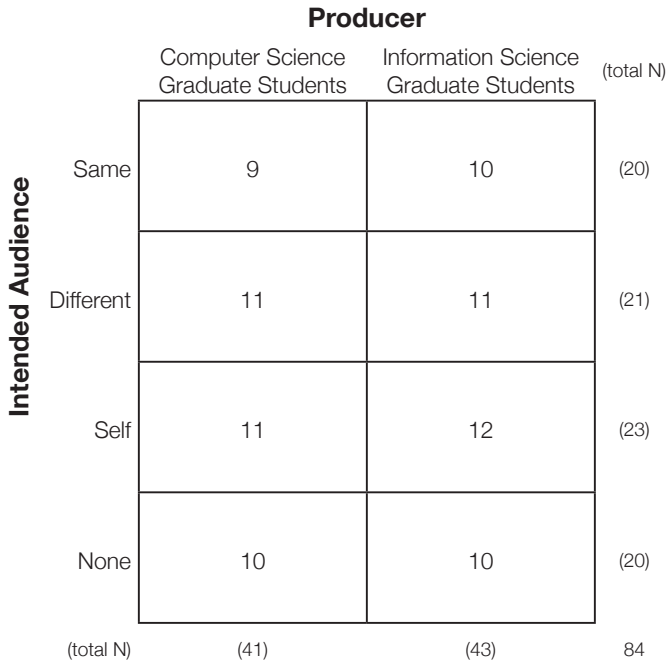


Figure 6.1 Conditions in the Organizing Phase of the experiment

6.2 Hierarchy Analysis

The goal of this analysis was to identify patterns in various measures of hierarchy characteristics that might help explain why the results of the Finding phase turned out the way they did. All of the measures and results are included in Appendix G; a selection are discussed in more detail below.

I analyzed the hierarchies along three broad dimensions: topology, vocabulary, and semantics.

- *Topology*: these measures capture some aspect of the surface structure of a hierarchy, like the number of folders, the average file depth, or the average number of children per folder. The word "topology" means both the topographical study of a physical place, or a network of interconnections like in an electric circuit or an intranet. I am using it in a similar sense here.
- *Vocabulary*: these measures focus on the specific vocabulary choices participants made when organizing. Vocabulary measures include things like how often participants chose the same words to refer to the same files, or the number of characters or words in a label.
- *Semantics*: neither Topology nor Vocabulary measures capture the essence of what makes a group of files "go together" in a folder. By "semantics" I mean the meaning

behind a grouping of files. This is impossible to measure directly in a quantitative way—only the participants know why they grouped certain files together. However, it is possible to measure whether the groupings of specific files made by one group of participants are different from those of another group.

Some of the measures I used were characteristics of the hierarchies I was able to measure directly, like recording a person's shoe size or height. The Topology measure of the number of folders is like this—I can look at a hierarchy (programmatically) and count how many folders it has, without needing to use another hierarchy for comparison.

Other measures I calculated were similarity measures, or what I am calling *pairwise comparison* measures. Some characteristics of the hierarchies only made sense when considered in the context of comparison to another hierarchy. For example, consider label agreement. I wanted to find out if participants in different conditions used different words to describe the same files—not a different number of words, different exact word choices. The only way this can be measured is by comparing hierarchies, and counting the number of words two people have in common for the same file; there is no absolute yardstick to use.

Most of the interesting Topology measures are surface features like shoe size in that they an be recorded directly, averaged, and compared across conditions. However, the interesting aspects of Vocabulary and Semantics only make sense when hierarchies are compared with each other. I will describe examples of each of the three kinds of measures below, along with results that have implications for understanding the *audience design* effect.

6.2.1 Topology Measure

Each hierarchy can be represented by a file-by-file matrix with 33^2 cells. To generate a file-by-file distance matrix, the number of steps between each pair of files in the hierarchy is calculated, and entered into the appropriate cell of the distance matrix. The $\mathrm{average.path.length}$ measure results from taking the average of all the values in the distance matrix.

The average number of steps from any file in a hierarchy to any other file can be used as an indication of the complexity of the hierarchy; for example, a hierarchy with files grouped into only a few folders at the same level has a lower $\mathrm{average.path.length}$ than a hierarchy with 4 or 5 levels and fewer files per folder. Figure 6.2 on page 73 represents a hierarchy from the experiment with a low $\mathrm{average.path.length}$ score, 3.15. The hierarchy in Figure 6.3 on page 74 has a high $\mathrm{average.path.length}$ score, 4.98. Figure 6.2 is an example of one of the least complex hierarchies in this experiment. Figure 6.3, on the other hand, is one of the most complex.

Figure 6.2 Hierarchy created by a CS participant for an IS *Imagined Audience*, with average.path.length = 3.15

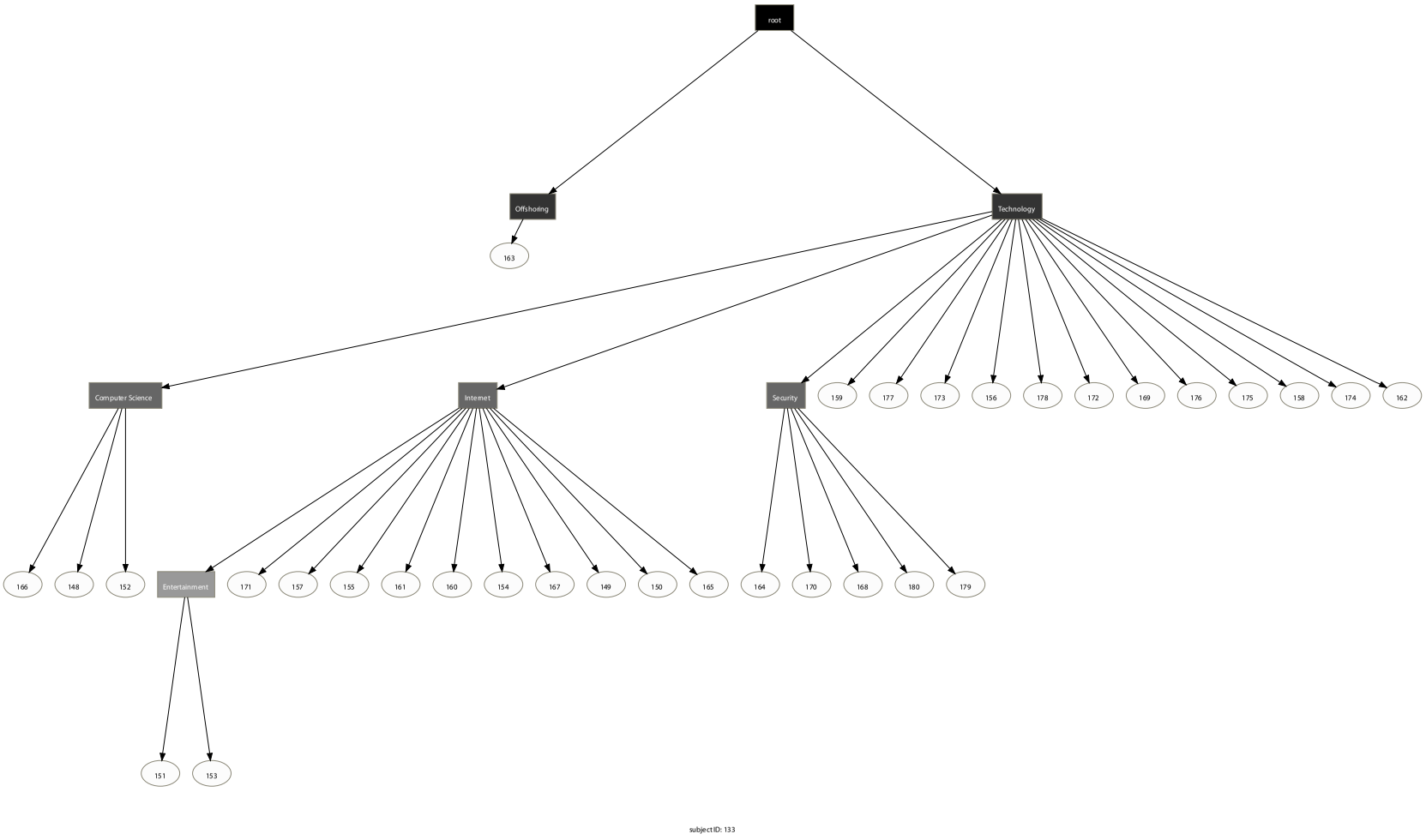
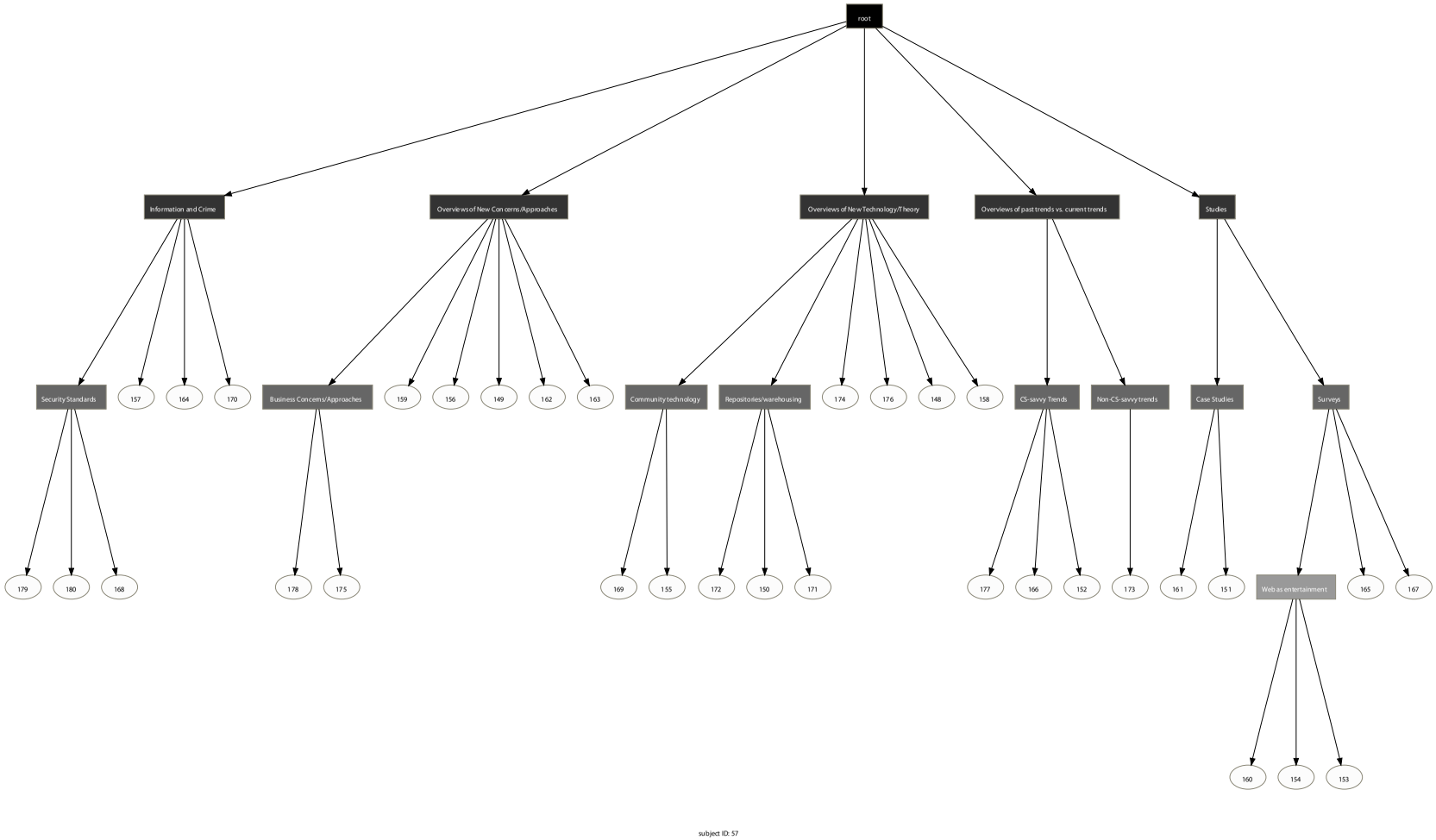


Figure 6.3 Hierarchy created by an IS participant for a CS *Imagined Audience*, with average.path.length = 4.98



I calculated the average.path.length for each hierarchy; the mean values for the levels of the *audience design* variable are presented in Table 6.1. A two-way *audience design* (Same, Different, None, Self) x *Producer community* (CS, IS) Analysis of Variance yielded no significant main effects or interactions.

Table 6.1 Mean average.path.length by *audience design* condition.

	M	*SD*	*N*
Same	4.00	0.67	19
Different	3.88	0.49	22
Self	3.86	0.50	23
None	3.83	0.51	20

In addition to average.path.length, I calculated a *pairwise comparison* Topology measure: file.adjacency. This measure is also based on a file-by-file matrix, but instead of differences, this matrix contains similarities. The cells of the file adjacency matrix are filled with a '1' if a pair of files in the hierarchy are grouped together in the same folder, and a '0' if they are in different folders. Two adjacency matrices are then compared, to yield a score representing how similar the two hierarchies are in terms of which files were grouped into the same folder. This measure does not represent files as grouped together if one is located in a subfolder of where the other file is stored.

To calculate file.adjacency, the pair of matrices that are being compared are "stacked up", and corresponding cells from each matrix are added together. In the matrix that results, the number of cells containing '2's is divided by the number of non-zero cells. This results in a score for a pair of hierarchies ranging from 0 to 1, representing the proportion of documents both users put together in the same folder. Two identical matrices, i.e., a hierarchy compared with itself, have a total score of 1.

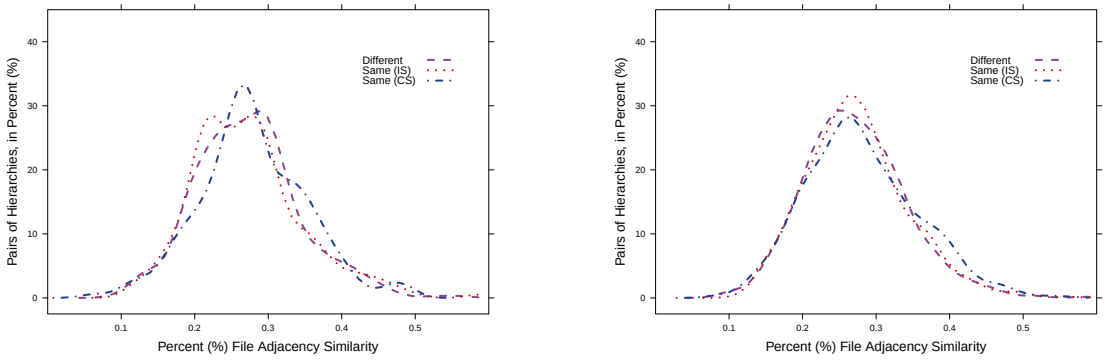
Figuring out how to calculate a *pairwise comparison* measure is just the first step; in order to use it, one must decide which pairwise comparisons are of interest. For example, I might compare all hierarchies that were made by people from the same *Producer* community, regardless of the *Imagined Audience* condition to which they were assigned. I could then calculate similarity scores for all possible comparisons of hierarchies within the CS *Producer* group, and within the IS *Producer* group, to find out if CS graduate students were more consistent with each other in the structure of their hierarchies than IS students.

However, in light of the result from the Finding phase of the experiment, it is more instructive to focus on comparisons that involve the levels of *audience design*. So, I created two subsets of the dataset, one containing just the hierarchies that were created for the same *Imagined Audience* that the *Producer* belonged to, and one with just the hierarchies created

for someone Different from the *Producer*. I then did pairwise comparisons between the hierarchies within each subset (Same vs. Different), and divided the scores from each subset into three groups so that I could calculate some descriptive statistics. The groups were based on whether the Producers of each hierarchy in the pairwise comparison were from the Same community (both CS vs. both IS), or from Different communities (one of each). So, for example, in one group all the producers were from the CS community (Same), and the pairs of hierarchies being compared were created for the Same *Imagined Audience*.

Returning to the Topology file.adjacency measure, as Figure 6.4 on page 76 shows there was very little difference in file.adjacency scores when comparing hierarchies created for someone from the producer's own community, vs. someone from a different community (Figures 6.5(a) vs. 6.5(b)). In other words, participants were not creating structurally different hierarchies based on the *Imagined Audience*. In all cases, the mean file.adjacency was very close to 0.27, meaning that regardless of target audience or producer community, participants tended to group the same files into the same folders about 27% of the time. The means are reported in Table 6.2. Kruskal-Wallis nonparemetric tests for main effects within each *Imagined Audience* group (Same and Different, the columns in Table 6.2) were not significant.

Figure 6.4 Histograms for the file.adjacency measure, comparing hierarchies created for others from the Same vs. Different communities, than the *Producer*. The histograms look very similar, indicating that the file.adjacency scores did not differ depending on whether the *Producer* and *Imagined Audience* were from the same community or not.



(a) *Imagined Audience*: Same

(b) *Imagined Audience*: Different

Neither the file.adjacency Topology measure, nor average.path.length showed meaningful differences that could help to explain the *audience design* effects observed in the Finding phase of the experiment. *Imagined Audience* seems to have had very little effect on the structural characteristics of the hierarchies created in the Organizing phase.

Table 6.2 Descriptive statistics for file.adjacency measure.

	***Imagined Audience*: Same**			***Imagined Audience*: Different**		
	M	*SD*	*N*	*M*	*SD*	*N*
Producers: Different	0.27	0.07	441	0.27	0.07	1322
Producers: Same CSE	0.28	0.07	382	0.28	0.08	629
Producers: Same MSI	0.27	0.08	422	0.27	0.07	692

6.2.2 Vocabulary Measures

The second dimension of the hierarchies that I analyzed was the Vocabulary dimension. Vocabulary measures all involve file and folder labels; for example, the number of words per label would be considered a Vocabulary measure. I will discuss two of these measures in this chapter; further examples of Vocabulary measures can be found in Appendix G.

mean.rank is a measure indicating how many unusual or unique words a participant has chosen to use in a file or folder label. To generate this measure, I followed the procedure of Krauss, Vivekananthan, and Weinheimer (1968) who asked participants to create names for a set of color chips. They found that participants chose more unique words when they were labeling color chips for themselves to find later, than when they were labeling for an unknown other person. I started by making a list of all the words from all of the file and folder labels in the experiment, sorted by how many times a word had been used (lowest to highest). Each word was then given a rank based on its frequency. The score for each hierarchy was the average of all the ranks for all the words in the file and folder labels.

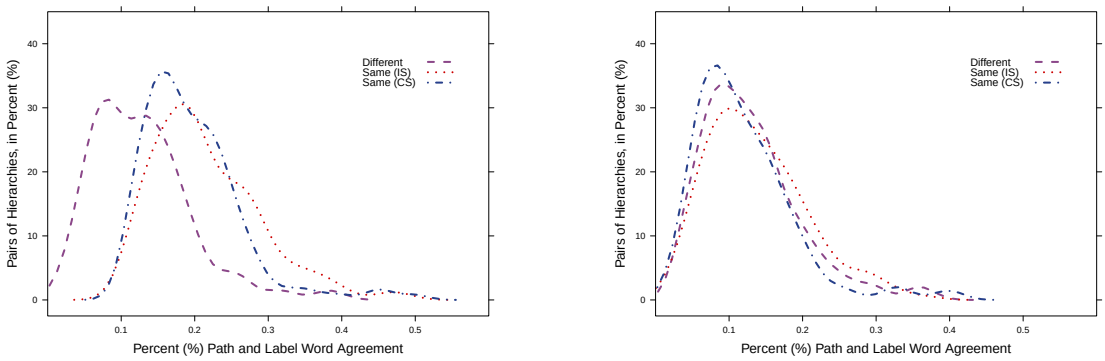
Table 6.3 mean.rank by *audience design* condition.

	M	*SD*	*N*
Same	39.08	4.97	19
Different	43.22	6.90	22
Self	42.93	6.51	23
None	38.95	4.74	20

There was a significant main effect for *Imagined Audience* ($F(3,76) = 3.39$, $p = 0.02$) in a two-way *audience design* (Same, Different, None, Self) x *Producer community* (CS, IS) ANOVA. The means for the levels of *Imagined Audience* can be found in Table 6.3; higher numbers mean more unique words were used. Overall, the pattern observed by Krauss et al. (1968) was also observed here; the mean for Different was the highest of the four *audience design* conditions, meaning participants organizing for someone different from themselves used more unique words.

I also computed a *pairwise comparison* Vocabulary measure: user.label.agreement. This measure represents the percent of the time two participants used the same exact words to label the same file. I used all words in the full path to a file as the "label" for that file, and did stemming and stop word removal before calculating user.label.agreement. When comparing two hierarchies, an user.label.agreement score was calculated for each file, and then averaged to create one score for the pair of hierarchies. The user.label.agreement scores for the *intended audience* conditions are listed in Table 6.4, and the histograms are displayed in Figure 6.5

Figure 6.5 Histograms for the user.label.agreement measure, comparing hierarchies created for others from the Same vs. Different communities, than the *Producer*. The histograms show that when the *Imagined Audience* community was Different from the *Producer* community, user.label.agreement is lower than when they were the Same. Also, in the left-hand graph, the user.label.agreement is lower when two hierarchies are compared that were created for different audiences.



(a) *Imagined Audience*: Same

(b) *Imagined Audience*: Different

The pattern of results in Figure 6.5 and Table 6.4 also indicates that something different was going on when a *Producer* labeled and organized files for someone who was not like them. Interuser agreement was in the range G. Furnas et al. (1983) observed in their series of studies on labeling (around 0.20) when the *Producer* and *Imagined Audience* were from the same community; however, user.label.agreement was quite a bit lower when they were not. The results for the two Vocabulary measures, mean.rank and user.label.agreement, both point toward differences in the selection of labels for files and folders, depending on whether or not the *Producer* was organizing for someone like themself.

Table 6.4 Descriptive statistics for user.label.agreement measure.

	***Imagined Audience*: Same**			***Imagined Audience*: Different**		
	M	*SD*	*N*	*M*	*SD*	*N*
Producers: Different	0.13	0.07	132	0.13	0.07	396
Producers: Same CSE	0.20	0.07	132	0.12	0.07	198
Producers: Same MSI	0.22	0.07	132	0.14	0.07	198

6.2.3 Analysis of Hierarchy Semantics

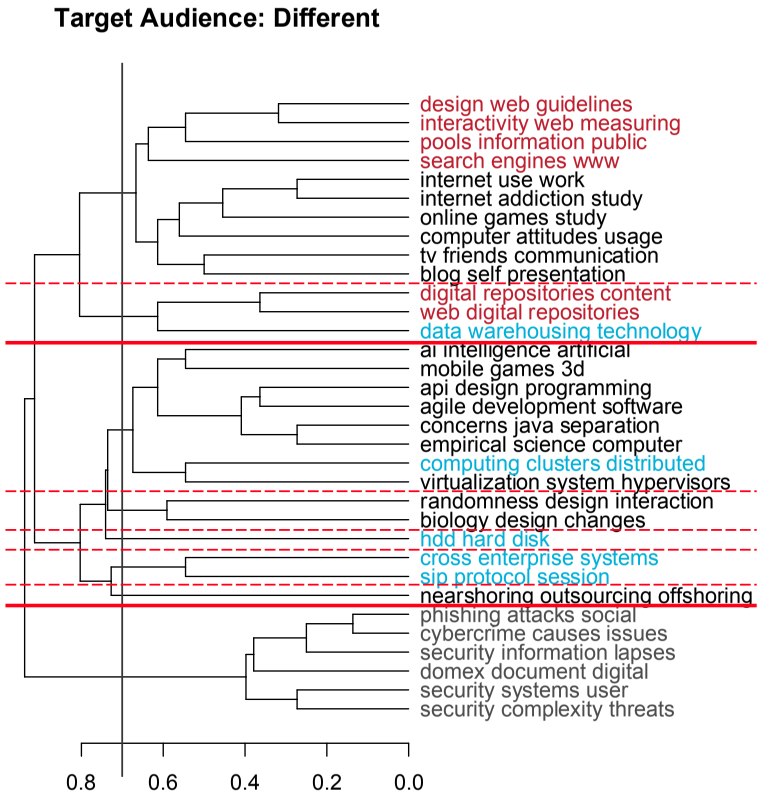
An important part of organizing files into a hierarchy is deciding which files "go together" in a given folder. Users tend to have mental rules that serve as both description of the contents of a folder, and filter for deciding what else makes sense to add to a folder (Whittaker & Sidner, 1996). The data I collected in the experiment do not support conclusions about what those mental rules might be, or how they might be different across *audience design* or *community membership* conditions in the experiment. However, the hierarchy data allows me to detect patterns in which files tended to be grouped with which other files, and there are several statistical techniques that allow me to visualize these similarities and differences.

To analyze the semantic groupings of files into folders, I used the file-by-file distance matrices created for the Topology analysis. But, instead of doing pairwise comparisons, I used them to do a hierarchical cluster analysis. I first divided up the hierarchies into four groups according to the *Imagined Audience* condition: Same, Different, Self, and None. I then "stacked up" the matrices within each group, like I did for the file.adjacency measure discussed above, and calculated an average for each cell of the matrix. The hierarchical clustering was created based on these average distance matrices, using the "complete" agglomeration method.

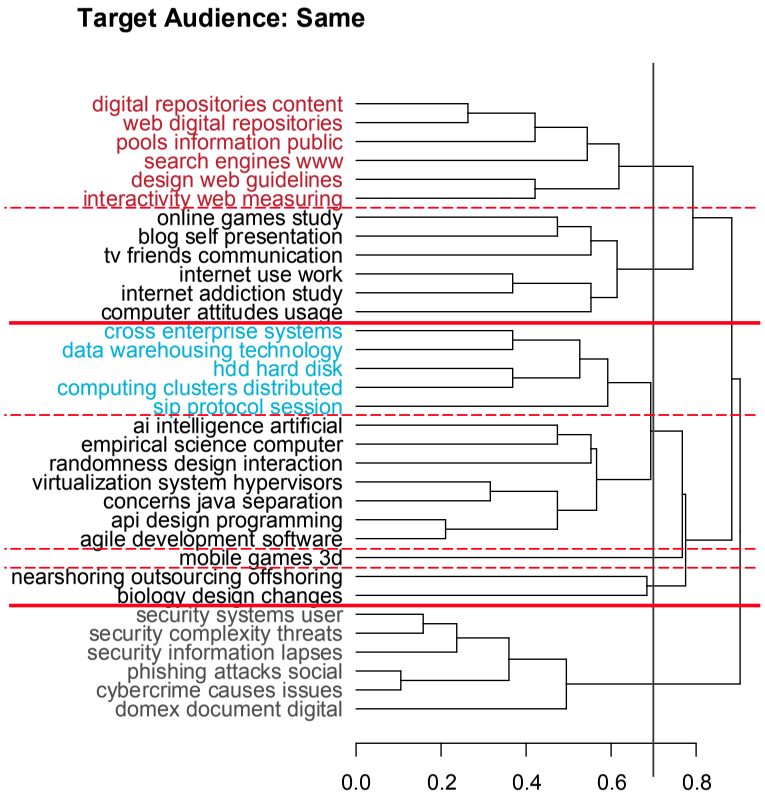
Hierarchical cluster analysis is a quantitative technique that provides diagrams that must be interpreted qualitatively. I used this technique to produce a series of cluster dendrograms that I then visually inspected for similarities and differences. This approach was motivated by findings in the categorization literature that experts and novices tend to categorize the same set of items differently, depending on their level of expertise (Ross & Medin, 2005). I wanted to see whether I could detect the same effect in the hierarchies, but depending on the *audience design* independent variable. Did participants group items differently, depending on who they were organizing for?

The dendrograms for the *Imagined Audience* conditions can be found on pages 80 and 81. First, some explanation of the colors and symbols in the charts: on each chart, I drew a dark line at approximately 0.7 on the horizontal axis; that line represents the cutoff point I used

Figure 6.6 Hierarchical cluster analysis dendrograms for the "Different" and "Same" conditions.

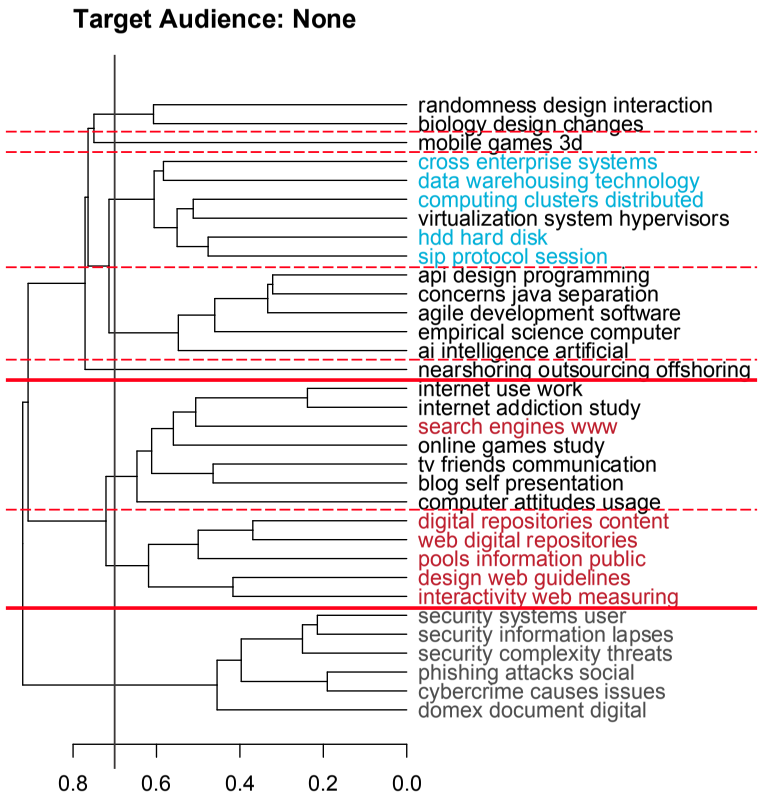


(a) Cluster analysis for the "Different" *Imagined Audience* condition

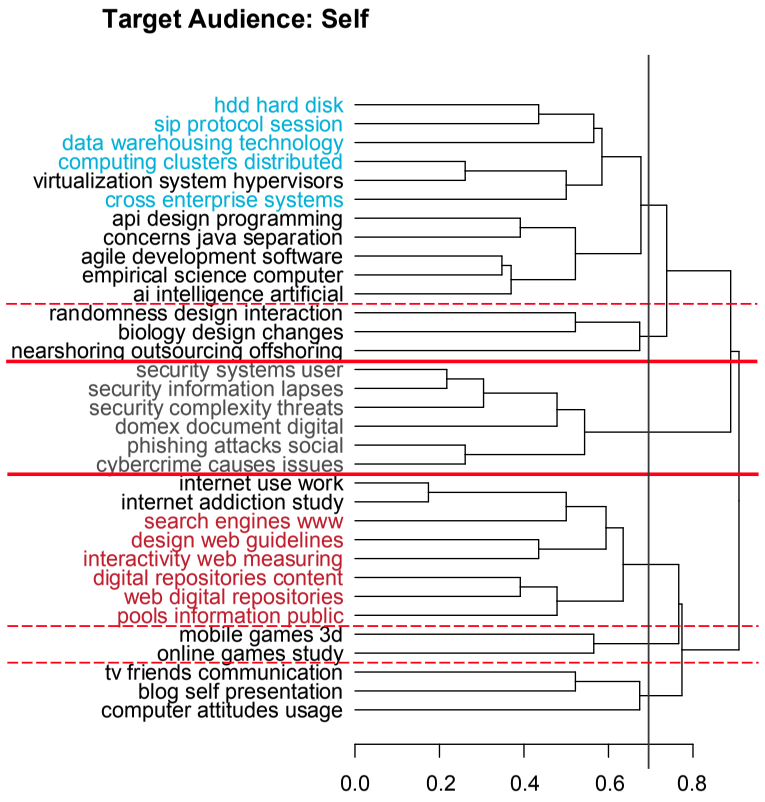


(b) Cluster analysis for the "Same" *Imagined Audience* condition

Figure 6.7 Hierarchical cluster analysis dendrograms for the "None" and "Self" conditions.



(a) "None" *Imagined Audience* condition



(b) Cluster analysis for the "Self" *Imagined Audience* condition

for identifying clusters. The solid red horizontal lines denote major cluster divisions, and the dashed red lines denote minor divisions within clusters. The list of files the participants organized is represented either on the left or the right side of the dendrogram—whether it is on the left or right is meaningless, I oriented them that way to make comparing the clusters across diagrams easier. These file "titles" are the three most common words participants used to label each file.

The "titles" are color-coded according to high-level topic. First, notice the titles that are gray in color—this cluster of files appears consistently in all four of the diagrams. These files are all related to computer security, and there was consensus among the participants (on average) that these files made up their own category or folder grouping.

The red "titles" are topically more relevant for the participants from the IS community, and the blue titles are more relevant for the CS community. I have colored the same files red and blue in each diagram, to illustrate how they were categorized differently by participants in different conditions. For example, start with the Same dendrogram (Figure 6.7(b) on page 80). The red files and the blue files are tightly clustered together. Looking at the Different dendrogram (Figure 6.7(a) on that same page), the clusters are no longer as tightly defined. In fact, in one sub-cluster, a blue file ("data warehousing technology") is grouped together with two red files. This increased variability between Different and Same is interesting in light of the *audience design* effect in the finding phase results. On the next page of diagrams (page 81) the dendrograms for Self and None are displayed. Neither shows the same degree of variability as the Different diagram. In the Self diagram, all the red files and blue files appear in the same cluster; in the None diagram there is one stray red file in a different cluster: "search engines www".

The cluster analysis hints at there being an influence of *audience design* in the way files are organized as well as named. It also suggests that future work should include a qualitative content analysis of the semantic structure of the hierarchies, so that these patterns can be explored further.

6.2.4 Hierarchy Analysis Summary

This section has described several approaches to analyzing the hierarchies created in the Organizing phase of the experiment. The Topology measures, or the measures of the surface features of the hierarchies, did not produce any interesting effects or significant differences that could help to explain the pattern of results from the Finding phase. However, two different Vocabulary measures showed effects of *audience design*, and both measures pointed toward greater label variability when Producers organized the files for someone from

a Different community. Cluster dendrograms illustrating differences in how participants grouped files together also pointed toward the idea that there was greater variability between participants in the Different *Imagined Audience* condition than in the Same condition.

These results are far from conclusive. However, the hierarchy analysis has provided a list of new measures that can be included in a regression model like the one introduced in Chapter 5. In the next section, I compare several new regression models that incorporate the hierarchy measures, to determine which model fits the data best, and which hierarchy measures seem to capture the same effects as the $\mathrm{PA.Same}$ variable in the original model.

6.3 Regression Model Comparison

In this section I use several hierarchy measures as predictors in a new series of regression equations for the Finding results. The goal of this analysis is to attempt to explain, based on measurable aspects of the hierarchies, how much these aspects contributed to the observed performance differences in the search tasks. The dependent variable of interest is still $\mathrm{total.clicks}$. However, in this section I compare a set of candidate models using Akaike's Information Criterion (AIC), (Anderson, 2008) to determine which model has the most explanatory power given the data and the models at hand.

6.3.1 Possible Predictors

The new hierarchy measure (predictors) are:

- $\mathrm{user.label.agreement}$: this is the same $\mathrm{user.label.agreement}$ measure as described in the previous section
- $\mathrm{dist.matrix.corr}$: this measure uses the file-by-file distance matrices from the previous section. For each search task, I calculated the correlation between the distance matrix of the hierarchy created by the participant completing the search tasks, and the hierarchy in which the participant is searching. This measure is a proxy for the semantic similarity between the Finding participant's own hierarchy, and the one being searched.
- $\mathrm{label.file.similarity}$: this measure is the correlation between the label of the search target file in the hierarchy that is being searched, and the text of the file itself. This measure is inspired by the "information scent" work of Pirolli (2005). Loosely stated, information scent is a property of hyperlink text—a cue that users pick up on that provides information which helps them make navigation decisions.

- path.file.similarity: this measure is identical to label.file.similarity, but instead of the correlation between the file label and the text of the search target, it is the correlation between the full path (without the label) and the text.

Controls introduced in the previous chapter, from page 58:

- shortest.path: for each search task, the depth in the hierarchy of the target file, i.e., the absolute minimum number of clicks, or "shortest path" to find the target
- average.path.length: the average number of steps from any file in a hierarchy to any other file, used as an indication of the complexity of the hierarchy; for example, a hierarchy with files grouped into only two folders at the same level has a lower average.path.length than a hierarchy with 4 or 5 levels and fewer files per folder
- consumer.id: because each person experienced all types of search tasks, the model includes a fixed effects control for individual differences

Finally, the experimental variables, also repeated from the previous chapter:

- imagined.audience: the *Imagined Audience* for whom the hierarchy was created
- PA.Same: are the *Producer* and *Imagined Audience* from the same community? Yes or No
- AC.Same: are the *Imagined Audience* and *Consumer* from the same community? Yes or No

6.3.2 Candidate Models

Based on the hierarchy analysis described above, there is evidence that the effect of PA.Same on total.clicks is due to differences in label choice between the two conditions. Therefore, I have included in the candidate models three measures of hierarchy vocabulary that are relevant for understanding behavior in the search tasks: user.label.agreement, label.file.similarity, and path.file.similarity, described above. In addition, the hierarchical cluster analysis hinted that organizing effects might have also played a role; I selected the dist.matrix.corr measure represent these effects.

I specified five poisson regression models, following the same assumptions as the model presented in Chapter 5 (see page 57). The goal of this analysis is to identify the model with the best fit to the data, given this candidate set of models, and to collect further evidence indicating whether differences in PA.Same resulted from *audience design* effects on participants' label choices.

The first candidate model, repeated here from the previous chapter, incorporates the independent variables from the experiment. It assumes $\mathrm{PA.Same}$ is important, and does not incorporate any of the new predictors:

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{imagined.audience},\\ &\mathrm{PA.Same}, \mathrm{AC.Same}, \text{imagined.audience} * \mathrm{AC.Same},\\ &\mathrm{PA.Same} * \text{AC.same}, \text{consumer.id})\end{aligned} \tag{6.1}$$

Alternatively, the next model (6.2), removes the experimentally manipulated variables, and replaces them with the four new hierarchy measures described earlier in the chapter:

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{dist.matrix.cor},\\ &\text{label.file.similarity}, \text{path.file.similarity}, \text{user.label.agreement},\\ &\text{consumer.id})\end{aligned} \tag{6.2}$$

The third candidate model replaces the two "information scent" indicators from the previous model, $\mathrm{label.file.similarity}$ and $\mathrm{path.file.similarity}$, with $\mathrm{PA.Same}$ only. A comparison between this model (6.3) and (6.2) will help to determine the relative importance of $\mathrm{PA.Same}$ and the "information scent" indicators:

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{dist.matrix.cor},\\ &\text{user.label.agreement}, \mathrm{PA.same}, \text{consumer.id},\end{aligned} \tag{6.3}$$

The fourth candidate model is a combination of (6.1) and (6.2) that incorporates two of the categorical independent variables from the experiment back into the model:

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{label.file.similarity},\\ &\text{path.file.similarity}, \text{user.label.agreement},\\ &\text{imagined.audience} * \mathrm{AC.Same}, \text{consumer.id})\end{aligned} \tag{6.4}$$

Finally, I included a fifth model as a baseline for goodness-of-fit comparisons. Model (6.5) below is atheoretical in that it includes one regressor for each participant, hierarchy, and document used in the experiment. These predictors represent the most likelihood I can account for given the design of the experiment. This model is useless for making theoretical

or practical generalizations, but can be used as a point of reference.

$$\log(\text{total.clicks}) = f(\text{hierarchy.id}, \text{consumer.id}, \text{target.id}) \tag{6.5}$$

6.3.3 Model Selection Process

To select the model with the best fit to the data, given a dataset and a set of candidate models, I have used an "information theoretic" approach based on Akaike's Information Criterion (AIC) (Akaike, 1981), which can be calculated from the maximized log-likelihood for each model, and used to compare them. This approach is loosely similar to dropping one predictor at a time from an OLS regression to assess its impact on the coefficients and R^2; however, unlike in OLS regression the only requirement here is that the exact same dataset is used for each case (i.e., the set of candidate models does not need to be nested). The choice of what predictors to include in each candidate model is motivated by the hypothesis the researcher is testing (Anderson, 2008; Burnham & Anderson, 2004).

Philosophically speaking, this approach assumes that the truth is out there, but any model we researchers can create is only an approximation of it. Further, our approximation is only as good as what we are able to measure and include in a model. The value of AIC is a representation of "the 'information' lost when a particular model is used to approximate full reality" (Anderson, 2008). Unlike R^2 in OLS regression, the AIC for a given model has no objective interpretation on its own. But when used to compare two models, the difference in AICs represents the difference in "information" lost, allowing the researcher to make a determination about which model is "closer" to truth. The goal is to select the model out of a candidate set that has the lowest AIC value.

An important caveat to this approach is that because we can never measure full reality, the dataset in question is taken as the best approximation available. If the data one has collected have little bearing on the phenomenon of interest, or if the constructs are poorly selected, or if measurement error exists (etc.), one can still produce an AIC value for each model and compare them. The argument about how close to "full reality" the selected model comes is based on the validity of the research design. This is why the comparison of AIC values is said to produce the best model, *given the dataset and the candidate models in question.*

Finally, this approach usually involves some assessment of the goodness-of-fit of the model to the data. If the candidate models are all "bad", it is still possible to choose the one with the lowest AIC value and claim that the "information lost" has been minimized.

Fit in OLS regression is represented by R^2—the proportion of variance 'explained' by the regression model. In maximum likelihood estimation (MLE), the goal is not to minimize the unexplained variance, it is to choose a set of coefficients (model parameters) that maximize the probability of producing the data given the model. There are many different ways to calculate "pseudo-R^2" for MLE models; however, they are not interchangeable. Magee (1990) wrote,

> "The statistical properties of these measures are largely unknown, and it seems unlikely that this knowledge would help in choosing one. They tend not to be desired for formal statistical reasons, but simply because researchers like to use the R^2 of the standard [OLS] regression model and would like to have something similar to report for other models." (p. 252)

I have chosen to use the $\bar{R}^2$ suggested by Nagelkerke (78) and endorsed by Anderson (2008), which can be interpreted as the proportion of the explained variation; however, because maximizing this value is not the goal, it should not be thought of as equivalent to the OLS R^2.

6.3.4 Results

Table 6.5 displays the model comparison results for the set of models introduced above. Because the number of observations relative the number of coefficients estimated in these models is relatively small, and because there is overdispersion in the dataset, an adjustment to AIC must be made (Anderson, 2008); $QAIC_c$ was used here instead of AIC. Δ_i is the difference between the $QAIC_c$ value for a given model and the minimum $QAIC_c$ in the set. w_i is the "Akaike weight"—it is calculated using Δ_i, and represents the probability that a given model is the best one given the comparison set and the data.

The rule of thumb when interpreting $QAIC_c$ given by Anderson (2008) is:

> "Models with delta values close to 0 have a lot of empirical support. Models with delta values in the rough range 4-7 have considerably less support, whereas models with delta values in the fringes (say 9-14) have relatively little support. Others, still further away, might be dismissed by most observers as implausible." (p. 85)

The model with the most empirical support in this set is (6.4); the second-best model is (6.1). Interestingly, the best model is identical to the second-best model, except for one difference: the second-best model uses the $\mathrm{PA.Same}$ experimental variable as a predictor, and the best model replaces $\mathrm{PA.Same}$ with three hierarchy Vocabulary measures:

Table 6.5 Model comparison table.

Rank	Model	QAIC_c	Δ_i	w_i	$\bar{R}^2$
2nd	(6.1) Experiment IV's Only	6618.53	7.08	0.028	0.23
3rd	(6.2) Hierarchy Measures Only	6628.05	16.60	0.000	0.24
4th	(6.3) IV's and Hierarchy Measures w/o Info Scent	6633.07	21.62	0.000	0.23
BEST	(6.4) All Variables	6611.45	0.00	0.972	0.25

label.file.similarity, path.file.similarity, and user.label.agreement. This is the first piece of evidence that the hierarchy differences captured by PA.Same in the model were related to label choice, and interpretation of those labels. The regression results for the best fit model are presented in Table 6.6.

The "pseudo-R^2", $\bar{R}^2$, for the best model is 0.25. To put this value in context, I calculated the goodness-of-fit for the athoretical model (6.5), on page 86. $\bar{R}^2$ for that model was 0.45—this is the best fit I could possibly achieve for these data given the variables available. QAIC_c for the athoretical model was 6341.45.

Table 6.6 Negative Binomial Regression estimates for the best fit model. Theta (dispersion parameter) = 2.779. consumer.id dummy variable coefficients are included in Appendix I. White's robust standard errors are reported.

	Regressors	*Estimates*	*Std. Error*	
0.	(Intercept)	0.747	0.238	**
1.	shortest.path	0.186	0.048	***
2.	average.path.length	0.239	0.055	***
3.	label.file.similarity	−0.320	0.134	*
4.	path.file.similarity	−0.677	0.188	***
5.	user.label.agreement	0.594	0.445	
6.	imagined.audience (Info. Sci.)	−0.053	0.129	
7.	imagined.audience (Self)	−0.042	0.107	
8.	AC.Same (Yes)	−0.056	0.131	
9.	imagined.audience (Info. Sci.) * AC.Same (Yes)	0.108	0.223	
10.	imagined.audience (Self) * AC.Same (Yes)	−0.143	0.182	

* $p < .05$; ** $p < .01$; *** $p < .001$

I performed one final analysis to collect additional evidence that could support or refute the idea that PA.Same-related effects were really due to label choices. One might argue that simply because I replaced one term with another in the regression equation is not enough to assume those predictors model the same phenomena. So, I performed four separate Kruskal-Wallis tests using the hierarchy measures label.file.similarity, path.file.similarity,

user.label.agreement and dist.matrix.corr as the dependent variables, and PA.Same as the independent variable. The purpose for these analyses was to determine whether these hierarchy-related scores differed, depending on whether the *Producer* organized and labeled from someone from their own community or not. The results of these tests are presented in Figure 6.8. Results showed significant differences for the three Vocabulary related measures (label.file.similarity, path.file.similarity, and user.label.agreement), but not for the Semantic measure (dist.matrix.corr). This indicates that participants made label choices differently, depending on the *audience design* condition.

Figure 6.8 Results of Kruskal-Wallis tests on four hierarchy measures, using PA.Same as the independent variable, showing that the hierarchies differed on vocabulary measures but not topography.

label.file.similarity
W (2, N = 1117) = 203.31, p < 0.0000

	M	*SD*	*N*
Same	0.4489	0.1962	372
Different	0.2679	0.1775	374
Self	0.2579	0.1498	371

path.file.similarity
W (2, N = 1117) = 64.20, p < 0.0000

	M	*SD*	*N*
Same	0.1175	0.1502	372
Different	0.0952	0.1552	374
Self	0.0504	0.0833	371

user.label.agreement
W (2, N = 1117) = 9.61, p < 0.008

	M	*SD*	*N*
Same	0.0957	0.0591	372
Different	0.0860	0.0548	374
Self	0.0848	0.0643	371

dist.matrix.corr
W (2, N = 1117) = 3.71, p < 0.16

	M	*SD*	*N*
Same	0.1819	0.1095	372
Different	0.1776	0.1118	374
Self	0.1701	0.1155	371

6.4 Discussion

In this chapter I presented three different dimensions of the hierarchies participants created in the Organizing phase of the experiment that can be operationalized and measured: topology, vocabulary, and semantics. I analyzed multiple measures of each type and found that while there were no meaningful differences in Topology when participants organized for different audiences, Vocabulary measures showed differences in the average number of unique words chosen, and the level of agreement between two participants in the labels they chose for the

same file, based on whether they organized for someone from their own community or not. Analysis of the Semantic component hinted at differences in the file groupings as well, and future qualitative content analysis will help to clarify these differences.

Model comparison among a candidate set of regression models using experimentally manipulated categorical variables and hierarchy measures as predictors yielded a "best model" that did not include $\mathrm{PA.Same}$, but instead included three hierarchy Vocabulary measures. This is evidence that the inclusion of the $\mathrm{PA.Same}$ predictor in the model captured some of the same patterns in the data as the Vocabulary measures. This idea is further supported by follow-up non-paremetric tests using $\mathrm{PA.Same}$ as the independent variable and the Vocabulary measures as dependent variables. These tests were significant, indicating that the categorical levels of $\mathrm{PA.Same}$ are related to differences in the hierarchy Vocabulary measures.

The next chapter will discuss the findings from the Organizing and Finding phases of the experiment, and the interview study, in the context of broader implications.

Chapter 7

General Discussion

7.1 Purpose and Research Questions, Revisited

The high-level goal of the research described in this thesis was to better understand how social factors contribute to the choices users make when organizing and labeling files in group information repositories, and how those choices affect users' finding behavior. In the first chapter, I introduced the diagram represented by Figure 7.1 as a starting point for my understanding of the relationship between information producers and consumers, and the repository. It depicts my argument that deciding what to call a file and where to put it in a repository are packaging choices made by information producers, and my hypothesis that these choices constrain where subsequent files can be stored and accumulate over time to create the structure within which information consumers try to find the files they need.

I asked the following research questions:

1. How do users share information using group information repositories? What influences their packaging choices, and how do they manage the information in the repository?

2. How do common ground and audience design affect file and folder labeling and organizing choices? How do these choices affect subsequent finding behavior?

These questions are fundamentally about users' interactions with shared artifacts, and whether the artifacts and the systems that house them can be considered both storage media and communications media. As such, this is applied research, focused on the interaction of the users with the technology, but grounded in theory from communications and psychology, and literature from computer-supported cooperative work.

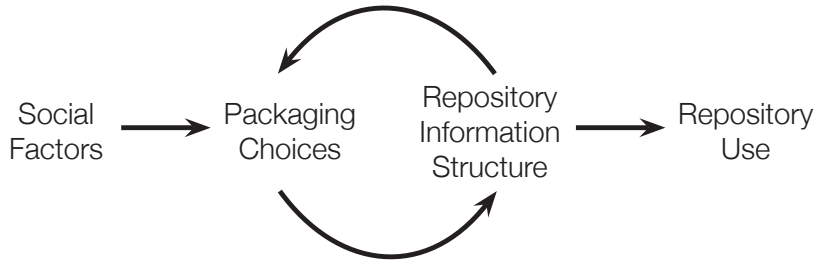


Figure 7.1 Diagram depicting the phenomena explored in this thesis.

7.2 Summary of the Results

7.2.1 Interviews and Log Data Analysis

Findings from the interview study indicate that social factors affect the information structure of the repository, and how it grows and evolves over time. Users restrict their activities to files they "own," are reluctant to delete files that might be useful to others, dislike the clutter that results, and can become demotivated if no one views files they uploaded. Group information repositories are different from tools for personal information management, and should be analyzed not just from the perspective of the functionality they embody and the information they contain, but also the social aspects of the context in which they are situated. This context influences the choices users make about where to store files, in essence constraining the way users package files in subtle ways that can be problematic as the site accumulates content over time.

From the interview study, I learned that social factors constrain users' choices in subtle ways that individually do not have much immediate impact on a repository. For example, users are hesitant to delete files because they do not want to prevent others from accessing them in the future. However, the choice NOT to delete, made for social reasons, means that repositories become more and more cluttered over time. Today's choice not to delete has consequences that are magnified in the future. I call this a "social" influence because users act out of concern for others, rather than out of a personal tendency to preserve or archive information, for example. Likewise, the finding that users of group information repositories stick to their own "turf" when choosing where to store files depicts a social phenomenon; the division of a space into territories seems to be a conflict-resolution strategy that is inherently social.

7.2.2 Organizing, Labeling, and Finding Experiment

The main finding of the experiment is that all participants searched more efficiently in hierarchies created by information producers who organized for an audience they believed to be similar to themselves, regardless of whether they were from the same community as the producer, or were actually in the target audience. In contrast, when producers tailored their hierarchies to dissimilar others, participants required the most clicks to reach the search target. Participants performed best when they were from the same community as the producer, were a member of the target audience, and the producer organized for "Self".

I analyzed the organizing data to better understand what aspects of the hierarchies created in the different experiment conditions led to the observed pattern of Finding results. I focused on three dimensions of the hierarchies participants created in the Organizing phase of the experiment: topology, vocabulary, and semantics, and found that while there were no meaningful differences in Topology when participants organized for different audiences, their labeling choices (Vocabulary) were affected by who they believed was the target audience of the hierarchy they were creating. Audience design had the expected effect when participants organized for themselves, and someone from the same community searched the hierarchy. However, when organizing for a community they were less familiar with, they seem to have made less sense.

From the experiment I learned that users can behave as though they are "communicating" through a group information repository. Information producers make an effort to tailor their organizing and labeling choices to the needs of others, and more importantly, these efforts affect later finding behavior. So once again, a choice made by a user today has enduring consequences for subsequent users of the repository. In addition, the experiment produced a surprising finding: when participants labeled files for someone they believed to be different from them, their attempt to tailor their choices for that audience did more harm than good—their label choices were more unique and made less sense. Participants had more difficulty finding files in hierarchies that had been created by someone who imagined their intended audience was from a different intellectual community.

With the experiment hypotheses I assumed that audience design would improve the communication between two interlocutors. In conversation, common ground accumulates with each utterance; however, in an asynchronous "not quite communication" medium like a group information repository, there is no support for the accumulation of common ground. One's model of the potential future intended audience is based on a priori beliefs and assumptions. Information that could be used to update or correct that model is not provided via group information repository systems. This means that any discrepancies between one's model of others' knowledge and the other person's actual knowledge are

perpetuated without hope of correction.

Because I did not experimentally manipulate or measure participants' mental models of the audience or their knowledge related to the content of the files, I cannot make any specific claims about these discrepancies. However, it is well established that in the absence of information to the contrary, people assume others know the same things they do, and experts overestimate the knowledge of others related to their own area of expertise (Nickerson, 2001). Perhaps the instructions of the experiment cause people to try harder than they otherwise would have to imagine what organization scheme might work best for the *Imagined Audience*, but without much information to rely on about that audience, the resulting hierarchies were more difficult to navigate.

7.3 Limitations

The goal of doing research is to learn something about the situation under scrutiny that one can then use to better understand the world. There are several limitations of this research that limit the implications and generalizability of this work.

Regarding the interview study, while I attempted to recruit as widely as was reasonable from the population of CTools project site users, the fact remains that CTools is just one system, and all participants are from the same institution. It is very difficult to make generalizable claims from just one case study. The experiment suffers from a similar sampling bias; neither study can be said to have a statistically random sample.

Regarding the experiment, there are certain findings for which I believe I have found a solid explanation. The link is clear between differences in label choices and audience design, and its effect on finding behavior. However, from this experiment I can only hypothesize why the labels were more variable when packaged for someone different from the participants; I did not talk to the users afterward about why they did what they did, or collect verbal protocols during the experiment.

I also cannot claim based on this experiment design that community membership common ground had no effect. I would need to do several things differently in order to conduct an experiment that would allow me to make such claims. For example, I would need a way to quantify and measure community membership common ground, and I would need to involve more than two communities such that the experiment involved participants having a wide spectrum of community membership common ground.

Another limitation has to do with the experiment task. In the real world, users of group information repositories do not organize a chunk of 33 documents all at once; instead, this

takes place gradually, one item at a time, over long timeframes. An attempt to conduct a controlled experiment over such a long timeframe would likely suffer from significant participant dropouts over time, and would be a much more costly and labor intensive effort. This experiment is a first attempt to find out whether common ground and intended audience have any effect in a highly controlled setting, and the findings will help in directing the next phase of the research.

I also might have included information or prompts in the interface to make the audience more or less salient to the participant while they were organizing, both in terms of making them more aware of their audience, and controlling what they know about that other person. In this experiment, I didn't measure how much participants knew about the intended audience, nor did I try to measure and manipulate what those mental models might have looked like. Because I didn't control for it or manipulate it, I can't make any claims from this research about how the model of others' knowledge might have affected packaging choices. For example, perhaps participants' mental models about their audience were wildly inaccurate when organizing for someone different from them, and that's what led to the variability in label choices. Because I didn't measure the mental models, I can't answer that question without doing additional research.

7.4 Implications

The high-level goal of this research was to gain insight into ways systems might incorporate better support for social aspects of group information management. I approached this goal by designing studies that allowed me to learn about the real-world use of repository systems, and also test hypotheses about the role of communication processes in labeling and organizing the contents of these systems. The qualitative findings showed me that users do in fact think about others when contributing, and the experimental results helped me to understand how this tendency to behave in a social way might actually do more harm than good when people need to find information in the system.

My definition of *packaging* has expanded through the course of this research to encompass more than the definition of Markus (2001). She referred to packaging as "the process of culling, cleaning and polishing, structuring, formatting, or indexing documents against a classification scheme" (p. 60). Based on my research, I now believe that any choice or action that affects the future use of a group information repository by another person can be considered packaging, and that these choices and actions happen all the time. Thinking about packaging as an everyday social activity for knowledge workers, and of repositories

as artifacts that are subject to social processes and constraints, is a shift in perspective from the idea of a repository as a static archive.

The key insight from this research is that these social factors not only affect users' choices and behavior—they also affect others' access to the information. Social forces shape the information infrastructure in lasting ways that can preclude—or facilitate—access to information. This makes group information repositories different from email or instant messaging, where the tool is constant no matter what information is conveyed. A repository is both acted upon by social forces, and can communicate information between people. As such, it is dynamic in ways other media are not. If it is viewed solely as a "storage" medium, access is about permissions and connectivity. If it is also viewed as a "social" artifact, access depends both upon the system and upon other people. So a shift in focus from "storage" to "social" means acknowledging the dynamic and evolving nature of the repository, recognizing that social forces are in important part of that evolution—and providing people with the functionality and information necessary to do a better job of collaboratively creating that artifact.

So what does it matter that these systems are social, and that one user's choices affects the next user's access? I began this thesis by introducing four examples of online information sharing and reuse:

- A scientist needs to locate some procedures and results from an experiment conducted by another researcher in his lab.
- A student learning the open-source, command-line statistical computing environment R needs to find out how to calculate the mode of her dataset.
- A new member of a design team needs to review requirements analysis activities that took place before he joined the team.
- An intelligence analyst needs to consult information collected by other agencies to assess a potential threat.

In each of these scenarios, the protagonist is looking for information that has been "packaged" by someone else. Knowing that packaging is a social process allows me to provide advice and suggestions to end users and system designers, so that users in these kind of situations have an easier time finding the information they need.

For individual users of repositories, it is important to remember that one's estimations about what labels and file groupings might work well for someone else are likely to be inaccurate, and these inaccuracies have measurable negative consequences for finding. Results of this research suggest that it is better to not get too crazy when organizing and labeling

for others; information scent (Pirolli, 2005) matters, and sticking close to the actual content when labeling and organizing is a better strategy than random speculation about what might work well for someone else.

Also, deleting files—or rather, not deleting files—seems to be an important socially motivated factor in making sites cluttered and unusable. Users should be more aware of their own tendencies not to delete, and ask rather than assume it is always better to keep information around indefinitely. System designers can create ways to help users make better packaging choices, too. For example, aggregate information about which files haven't been looked at in a really long time might make the task of deciding what to delete, or even having a discussion about likely candidates, much easier.

Finally, the fact that group information repositories are considered to be a central location for group members to store and organize their information supports certain assumptions and expectations that everyone has a similar conceptualization of and level of familiarity with the contents of the system. However, in practice these assumptions are invariably incorrect. This thesis highlights a breakdown in the flow of information from producer to consumer, between naming/organizing and finding; the content of the files are conveyed, but information about the people is lost. This means that when somebody has trouble finding a file they need, unlike misunderstandings in face-to-face conversation, there is no way for them to attempt to negotiate shared meaning and repair the disconnect. What producers and consumers know about each other usually comes from sources outside the system; but, if users were given the right information and feedback at the right time, they might be able to communicate better *through* the system.

I suspect that finding ways to incorporate support for the formation of more accurate mental models of other users' knowledge could help users package more effectively. Krauss and Fussell (1991) conducted a series of studies to investigate perspective-taking in conversation, or "the processes by which people estimate what others know" (p. 11). They discovered that people tend to overestimate the prevalence of their own knowledge in others; in other words, we all at least initially believe other people know about the same things we do. It is only after some interaction takes place that we learn otherwise and adjust accordingly. Krauss and Fussell found that when they did not allow their participants to interact with each other, perspective-taking was very difficult, and initial biases persisted much longer than when participants were able to communicate directly.

User behavior can be influenced merely by information included in the user interface (Sen et al., 2006; Lampe, Ellison, & Steinfield, 2007; Shami, Ehrlich, Gay, & Hancock, 2009). These results hint that incorporating information that makes the audience more salient and familiar could help users form better mental models of other users, and find the

information they need in group information systems. Determining what the right information might be, and how it should be presented to the user, are left for future work.

7.5 Future Directions

First, I want to look for evidence of packaging (audience design) in other real-world systems that support information sharing. In the experiment, simply providing different instructions to participants was enough to induce them to package differently; this is evidence that information producers need very little information to tailor their information sharing contributions for consumers. However, packaging may take place to varying degrees in different systems and at different scales. Identifying instances of packaging will provide a foundation for understanding what design approaches are most effective.

Second, it is not clear how aware producers and consumers are of each other when they use information sharing systems. In email or instant messaging, the participants in the conversation are known to each other. However, the communication participants are less well-defined when contributing Facebook status updates or Tweets, tagging, or posting photos. Who do producers and consumers feel like they are "talking to", if anyone? How do these perceptions affect what people choose to share, and how the information is packaged? Understanding these perceptions, and what triggers them, will help designers make choices that appropriately support awareness between producers and consumers.

Third, if producers and consumers are aware of each other when using information sharing systems, it is likely they use whatever information is available to them to make inferences about each other. Users' perceptions of each other are important, because they provide the starting point for the negotiation of shared meaning in these contexts. However, this aspect of information sharing has received little attention. What information do consumers need about producers to make sense of their contributions more effectively? When producers contribute information, how does what they know about their intended audience affect how the information is packaged? What information will help users form better mental models of each other, and how and when should that information be provided to users to influence their packaging choices?

Finally, the unexpected results of the experiment highlight an opportunity to explore the theory of common ground and the grounding process as it relates to shared artifacts—not in the sense of referential communication tasks or shared visual information (Gergle et al., 2006), but rather situations in which common ground is mediated by an artifact, like a repository. Grounding, as it is defined by Clark and Brennan (1991) depends heavily on

real-time visual and auditory feedback; perhaps a different process altogether is at work when people create shared understanding both about and through a “thing”.

Appendices

Appendix A

Interview Study Protocol

Warm-Up

- Introduce myself, purpose of study, how I selected the site and what I know about it (only the title), what Ill be asking about (questions about the ctools site and the project, tour of the contents of the site, how the information is used)
- Tell me a little bit about yourself and the work you do...

Grand Tour

- Tell me about the site – whats the purpose? Whats your role? How long has it been around?
- Tell me about how you use the site... how often, for what kinds of things? What kinds of items are in the site? Who are the other primary users?
- Describe the structure of the site (from memory, without looking at it)?
- Lets open it up now... Can you give me a tour of the site talk about each folder and what is in it?

Files Used, Added, Deleted

- When was the last time you used the repository before this study session? Tell me about that time – what were you doing – why did you need the file?
- Can you find that file for me now (think out loud)? When was the time before that?
- Can you remember a time that you had trouble finding something? Tell me about it... who added/named the file? Where did you expect to find it? Where actually was it? What do you think of the name? When was the time before that?
- Which other files do you remember using? Tell me more about how you used this particular file what was the situation? Why did you need the file? (Probe for who created it, how old it is, how recently it has been used, the story of how they used it.)

- What files do you use that you know somebody else put up there?
- What files do you think are accessed the most? Say more about that...
- Tell me about “Notifications”... do you use them? What was the last notification you received? Sent? What about “Announcements”?
- Which files do you remember adding? When was the last time you added a file? Tell me more about when you added that file what was the situation? Why did you add the file? (Probe for information about labeling and organizing choices and decisions.)
- When was the last time you deleted something? etc.
- Can you remember a time when somebody else used a file that you added or use frequently? How did you know they were using it (did they ask for it, etc.)?

Site Organization

- Can you think of a file that is an important file, or used often by the group? Show me. Open the file, tell me about it... Another file? What parts of the site get a lot of use? Not very much use?
- Has any member of the site ever “cleaned up” the files and folders? Say more about that...
- Are there any naming or organizing conventions you can think of for this CTools site? What explicit guidelines, rules, agreements exist for where things are stored and how they should be named?

Wrap-up

- Does your group have other files that are kept elsewhere?
- What are other ways your group shares files, etc.?
- How has the site evolved in the time youve been using it?
- How would you describe the repository to a newcomer to your group – what would be important for them to know?

Appendix B

Instructions for the Online Experiment, Organizing Phase

B.1 Organizing Phase Overview

Participants labeled and organized documents into folders, using an online interface that designed to resemble the traditional file-and-folder interface that is prevalent in most personal computer operating systems, as well as most group information repository systems. Characteristics of this type of interface include the ability to rename files, and drag and drop files into mutually exclusive folders (i.e., each document can exist in only one place). The sequence of activities for this experiment was as follows:

1. Potential participants responded to recruiting advertisement. The experimenter verified that they were students in the appropriate department, in the UM Directory (http://directory.umich.edu).

2. Experimenters contacted eligible respondents by email, to provide more information about the experiment.

3. If participants still expressed interest, they received a username and password for the experiment system.

4. View login page and online consent form. If participants clicked "I Agree" they were allowed to proceed to the experiment.

5. View instructions for the experiment, and a practice task so participants could become familiar with the experiment interface.

6. View additional instructions, and complete the Labeling and Organizing task.

7. Complete a series of questionnaire items. Participants were then asked for an email address for delivery of the study incentive, a $15 Amazon.com gift card. Participants were also informed about Part 2 of the experiment, and reminded that they might be contacted again.

B.2 System Functionality and Tutorial Instructions

This experiment is Part 1 of a two-part experiment. Part 1, which you will be completing today, consists of three different kinds of tasks. In the first task, you will create labels for documents, and organize them into folders. In the second, you will be asked to describe your mental "rules" for each folder you've created. In the third task, you'll answer a series of questions about yourself, the interface, and the organizing task. Please set aside enough time (about 60 min) to complete all three tasks in the experiment in one sitting.

The file labeling and folder creation interface works a lot like the file-and-folder interface you're used to using in Microsoft Windows or Mac OS. In this experiment, you begin with a list of files that have very generic labels. Clicking once on a file's name allows you to view it; click a second time and you can rename the file.

A **New Folder** button in the interface allows you to create a new folder and give it a name. Renaming folders works the same way as renaming files. You can drag and drop files from the **Unfiled Documents** list into folders. You can also drag and drop folders into other folders, to create levels of sub-folders.

In the interface, documents and folders initially appear in alphabetical order. You can rename files and folders as many times as you want. The hierarchy you create can have as many levels as you think are appropriate, and folders can contain any number of files. Each file can exist in one and only one place; this system does not support aliasing, metadata, or multiple categorization.

A few notes about viewing and interacting with the experiment:

- it looks best on a screen resolution set to 1024x768 or better
- if a document doesn't display properly, simply reload the page and this will correct itself
- a blue progress bar at the top of the screen will update periodically to let you know how much of the experiment remains to be completed

If your web browser crashes or the browser window is accidentally closed, don't panic! You can return to the experiment using the link that was emailed to you, and when you log in you'll pick up right where you left off. If you encounter bizarre technical problems of any kind, please email Emilee Rader at ejrader@umich.edu.

In order to receive the $15 Amazon.com gift certificate for completing Part 1 of the experiment, you must make an honest attempt to complete all of the tasks and answer the required survey questions. This means, for example, that if you do not create meaningful labels for the documents and simply organize them all into a single "miscellaneous" folder, you will not be eligible for the gift certificate.

Participating in Part 1 of this experiment means that you will be eligible for Part 2, which will take place in about a month. More information about Part 2 will be provided as the experiment progresses.

Click the OK button below to proceed.

B.3 Practice Organizing Task Instructions

This is a practice task so you can get used to how the experiment interface works. Change the labels on the files (for example, recipe01) to something more meaningful, and organize them into folders. Click the "Done Organizing" button when you are finished; it will become active when all documents have been moved from the "unfiled documents" list. Please use English when creating names for files and folders.

B.4 Organizing Task: Scenario and Instructions

B.4.1 Instructions: no audience

Now that you've had a chance to become familiar with the interface, you are ready for the next step. On the following screen, you will be presented with a list of files. Each one contains a short article summary or excerpt. You may or may not already be familiar with the topics and concepts in the files. Your task is to create a more descriptive label for each file, and organize the files into folders. When thinking about what to name the files and what

folders to put them in, imagine that you are working on writing a literature review paper.

There are many different ways to go about completing this task. Some people prefer to read through all of the files and create labels, before organizing them into folders. Others label a few at a time and create folders as they go, renaming and rearranging folders as necessary. What process to follow is completely up to you.

Click the OK button below to begin.

B.4.2 Instructions: for yourself

Now that you've had a chance to become familiar with the interface, you are ready for the next step. On the following screen, you will be presented with a list of files. Each one contains a short article summary or excerpt. You may or may not already be familiar with the topics and concepts in the files. Your task is to create a more descriptive label for each file, and organize the files into folders.

There are many different ways to go about completing this task. Some people prefer to read through all of the files and create labels, before organizing them into folders. Others label a few at a time and create folders as they go, renaming and rearranging folders as necessary. What process to follow is completely up to you.

When thinking about what to name the files and what folders to put them in, imagine that you are working on writing a literature review paper, and you will need to find some of the files later.

In fact, you may be invited back to participate in Part 2 of this research study, in which you may actually be asked to find files in the hierarchy you created in this part of the experiment. So, when you are thinking about how to organize the files, please focus on creating a hierarchy with an organizational structure that makes the most sense for you.

Click the OK button below to begin.

B.4.3 Instructions: for IS students

Now that you've had a chance to become familiar with the interface, you are ready for the next step. On the following screen, you will be presented with a list of files. Each one contains a short article summary or excerpt. You may or may not already be familiar with

the topics and concepts in the files. Your task is to create a more descriptive label for each file, and organize the files into folders.

There are many different ways to go about completing this task. Some people prefer to read through all of the files and create labels, before organizing them into folders. Others label a few at a time and create folders as they go, renaming and rearranging folders as necessary. What process to follow is completely up to you.

When thinking about what to name the files and what folders to put them in, imagine that you are working on writing a literature review paper for a group project with Masters students from the School of Information (MIS) at the University of Michigan, and other members of your group will need to find some of the files later.

In fact, MIS students will be invited to participate in Part 2 of this research study, and they may actually be asked to find files in the hierarchy you will be creating in this part of the experiment. So, please focus on creating a hierarchy with an organizational structure that would make the most sense for MIS students.

Click the OK button below to begin.

B.4.4 Instructions: for CS students

Now that you've had a chance to become familiar with the interface, you are ready for the next step. On the following screen, you will be presented with a list of files. Each one contains a short article summary or excerpt. You may or may not already be familiar with the topics and concepts in the files. Your task is to create a more descriptive label for each file, and organize the files into folders.

There are many different ways to go about completing this task. Some people prefer to read through all of the files and create labels, before organizing them into folders. Others label a few at a time and create folders as they go, renaming and rearranging folders as necessary. What process to follow is completely up to you.

When thinking about what to name the files and what folders to put them in, imagine that you are working on writing a literature review paper for a group project with graduate students in Computer Science and Engineering (CS) at the University of Michigan, and other members of your group will need to find some of the files later.

In fact, graduate students in CS will be invited to participate in Part 2 of this research study,

and they may actually be asked to find files in the hierarchy you will be creating in this part of the experiment. So, please focus on creating a hierarchy with an organizational structure that would make the most sense for graduate students in CS.

Click the OK button below to begin.

B.5 Incentive and Contact Information Instructions

You have completed Part 1 of the Organizing and Finding Files Experiment. To thank you for your time and effort, you will receive an electronic gift certificate from Amazon.com, worth $15. You will be contacted again in approximately one month about participating in Part 2 of the experiment. Participants in Part 2 will receive an additional $20 Amazon.com gift certificate.

Please enter an email address where you prefer to receive information from Amazon.com about your gift certificate. The contact information you provide here will ONLY be used to send you the gift certificate, and to invite you to participate in Part 2 of the experiment; it will not be associated with your data, or used to identify your responses in any way. See the privacy notice on Amazon.com if you have any concerns about how they may use your email address. [link to privacy notice: http://www.amazon.com/gp/help/customer/display.html/ref=g_gc-dp_tc_p?ie=UTF8&nodeId=468496 – opens in new window]

You may choose to provide some other form of contact information (postal address, telephone number, etc.); however, if you do not provide an email address, it will take longer to arrange with you how you'll receive the gift certificate.

Enter your email address or other contact information below:
text entry box

Do you want to receive an email with the correct answers to the Analogy Questions, and your score?
yes, please; no thank you

B.6 Closing Screen Text

Thank you for participating! Please contact Emilee Rader at ejrader@umich.edu if you do not receive your gift certificate within two weeks, or if you have any questions or concerns about this experiment.

Appendix C

Instructions for the Online Experiment, Finding Phase

C.1 Finding Phase Overview

In the Finding Experiment, participants were presented with a series of hierarchies paired with a search target document. They were instructed to find the target document in the hierarchy. When one search task was completed, the next task appeared automatically. The sequence of activities was as follows:

1. Participants from Part 1 of the experiment were contacted approximately four to six weeks after completing Part 1, and invited to participate in Part 2.
2. When participants agreed to participate, they were sent login information for the experiment system.
3. View login page and online consent form. If participants clicked "I Agree" they were allowed to enter the experiment.
4. View nstructions for the experiment, and complete a practice task so participants could become familiar with the experiment system.
5. View sdditional instructions, and complete the Finding task.
6. When complete, participants were asked for an email address for delivery of the study incentive, a $20 Amazon.com gift card.

C.2 System Functionality and Tutorial Instructions

This is Part 2 of a two-part experiment. In this experiment, you'll be searching for a series of documents in file-and-folder hierarchies that were created by participants in Part 1 of the experiment. You will be shown a document on the right hand side of the screen, and a hierarchy on the left. Your task is to find the document in the hierarchy. When you have completed all the search tasks, you will be asked to answer a set of survey questions.

How the interface works:

- You can browse the hierarchy on the left side of the screen; it works a lot like the file-and-folder interface you're used to using in Microsoft Windows or Mac OS. There is no full-text search feature.
- The document viewing pane on the right side of the screen shows the document you are to find.
- You can click on documents in the hierarchy; they will open in another tab in the document viewing pane.
- When you've found the target document (denoted by the "Find This Document" tab), drag it from the hierarchy onto the "drop found document here" area at the bottom of the screen (surrounded by a dotted line) and drop it there.
- The next target and hierarchy will appear automatically. this will continue until all the search tasks are completed.

Note: the interface will NOT tell you whether or not you've found the right document, and you cannot go back and change your mind once you have dragged a document onto the "Drop Found Document Here" area.

This is not a timed task; take all the time you need. Please set aside enough time (about 60 min) to complete the entire experiment in one sitting.

A few notes about viewing and interacting with the experiment:

- It looks best on a screen resolution set to 1024x768 or better.
- If a document doesn't display properly, simply reload the page and this will correct itself.
- A blue progress bar at the top of the screen will update periodically to let you know how much of the experiment remains to be completed.

If your web browser crashes or the browser window is accidentally closed, don't panic! You can return to the experiment using the link that was emailed to you, and when you log in you'll pick up right where you left off. If you encounter bizarre technical problems of any kind, please email Emilee Rader at ejrader@umich.edu.

In order to receive the $20 Amazon.com gift certificate for completing Part 2 of the experiment, you must make an honest attempt to complete all of the finding tasks and answer the required survey questions.

Click the OK button below to proceed.

C.3 Practice Finding Tasks Instructions

This is a practice task so you can get used to how the experiment interface works. Find the target document in the hierarchy. When you've found it, drag the file from the hierarchy onto the "Drop Found Document Here" area at the bottom of the screen. The next search target and hierarchy will appear automatically. The blue progress bar at the top of the screen tells you how much of the experiment is left for you to complete.

C.4 Finding Task Instructions

Now that you've had a chance to become familiar with the interface, you are ready for the next step. You will be shown a document on the right hand side of the screen, and a hierarchy on the left. Your task is to find the document in the hierarchy. Clicking on documents in the hierarchy will cause them to display in new tabs on the right hand side of the screen, just like in the organizing task.

When you've found the document, drag the file from the hierarchy onto the "Drop Found Document Here" area at the bottom of the screen. The next search target and hierarchy will appear automatically. The blue progress bar at the top of the screen tells you how much of the experiment you have completed.

The interface will NOT tell you whether or not you've found the right document, and you cannot go back and change your mind once you have dragged a document onto the "Drop Found Document Here" area.

Click the OK button below to proceed.

C.5 Incentive and Contact Information Instructions

You have completed Part 2 of the Organizing and Finding Files Experiment; the experiment is now finished. To thank you for your time and effort, you will receive an electronic gift certificate from Amazon.com, worth $20.

Please enter an email address where you prefer to receive information from Amazon.com about your gift certificate. The contact information you provide here will ONLY be used to send you the gift certificate; it will not be associated with your data, or used to identify your responses in any way. See the privacy notice on Amazon.com if you have any concerns about how they may use your email address. [link to privacy notice: http://www.amazon.com/gp/help/customer/display.html/ref=g_gc-dp_tc_p?ie=UTF8&nodeId=468496 – opens in new window]

You may choose to provide some other form of contact information (postal address, telephone number, etc.); however, if you do not provide an email address, it will take longer to arrange with you how you'll receive the gift certificate.

Enter your email address or other contact information below:
text entry box

Re-enter your contact information:
text entry box

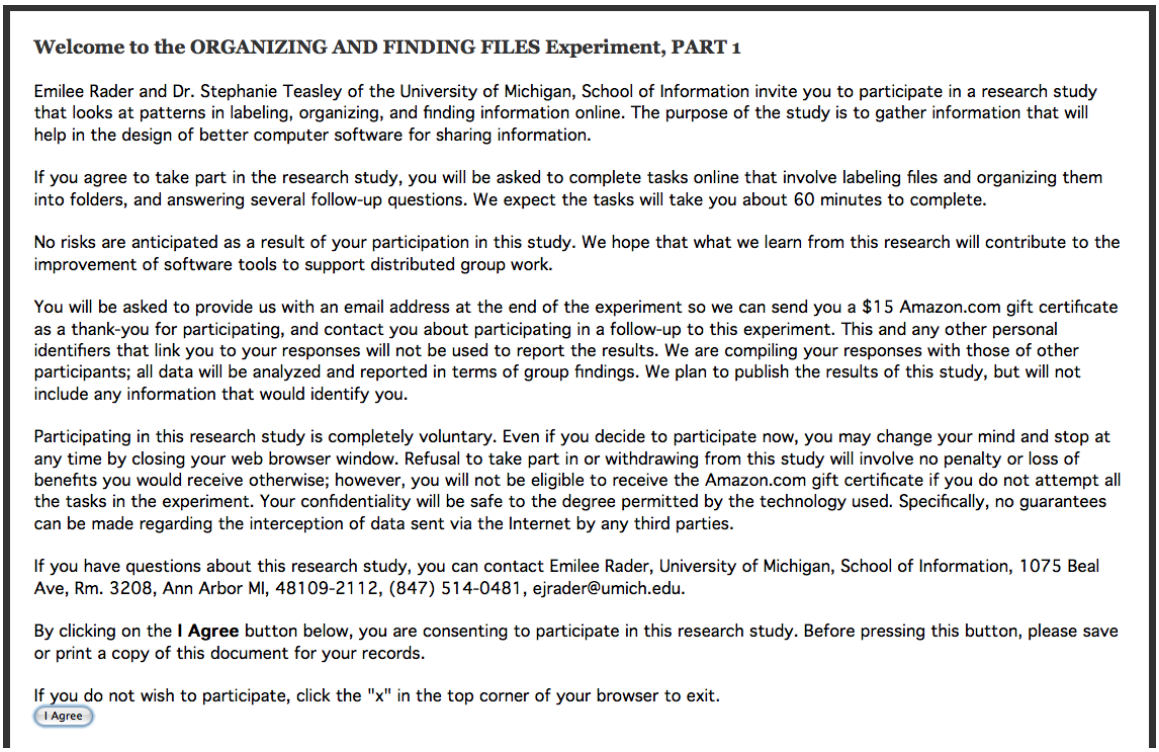
C.6 Closing Screen Text

Thank you for participating! Please contact Emilee Rader at ejrader@umich.edu if you do not receive your gift certificate within two weeks, or if you have any questions or concerns about this experiment.

Appendix D

Screen Captures from the Experiment Application Interface

Below are four screen captures depicting the user interface of the application created for the Online Experiment.



Welcome to the ORGANIZING AND FINDING FILES Experiment, PART 1

Emilee Rader and Dr. Stephanie Teasley of the University of Michigan, School of Information invite you to participate in a research study that looks at patterns in labeling, organizing, and finding information online. The purpose of the study is to gather information that will help in the design of better computer software for sharing information.

If you agree to take part in the research study, you will be asked to complete tasks online that involve labeling files and organizing them into folders, and answering several follow-up questions. We expect the tasks will take you about 60 minutes to complete.

No risks are anticipated as a result of your participation in this study. We hope that what we learn from this research will contribute to the improvement of software tools to support distributed group work.

You will be asked to provide us with an email address at the end of the experiment so we can send you a $15 Amazon.com gift certificate as a thank-you for participating, and contact you about participating in a follow-up to this experiment. This and any other personal identifiers that link you to your responses will not be used to report the results. We are compiling your responses with those of other participants; all data will be analyzed and reported in terms of group findings. We plan to publish the results of this study, but will not include any information that would identify you.

Participating in this research study is completely voluntary. Even if you decide to participate now, you may change your mind and stop at any time by closing your web browser window. Refusal to take part in or withdrawing from this study will involve no penalty or loss of benefits you would receive otherwise; however, you will not be eligible to receive the Amazon.com gift certificate if you do not attempt all the tasks in the experiment. Your confidentiality will be safe to the degree permitted by the technology used. Specifically, no guarantees can be made regarding the interception of data sent via the Internet by any third parties.

If you have questions about this research study, you can contact Emilee Rader, University of Michigan, School of Information, 1075 Beal Ave, Rm. 3208, Ann Arbor MI, 48109-2112, (847) 514-0481, ejrader@umich.edu.

By clicking on the **I Agree** button below, you are consenting to participate in this research study. Before pressing this button, please save or print a copy of this document for your records.

If you do not wish to participate, click the "x" in the top corner of your browser to exit.

I Agree

Figure D.1 Consent Form

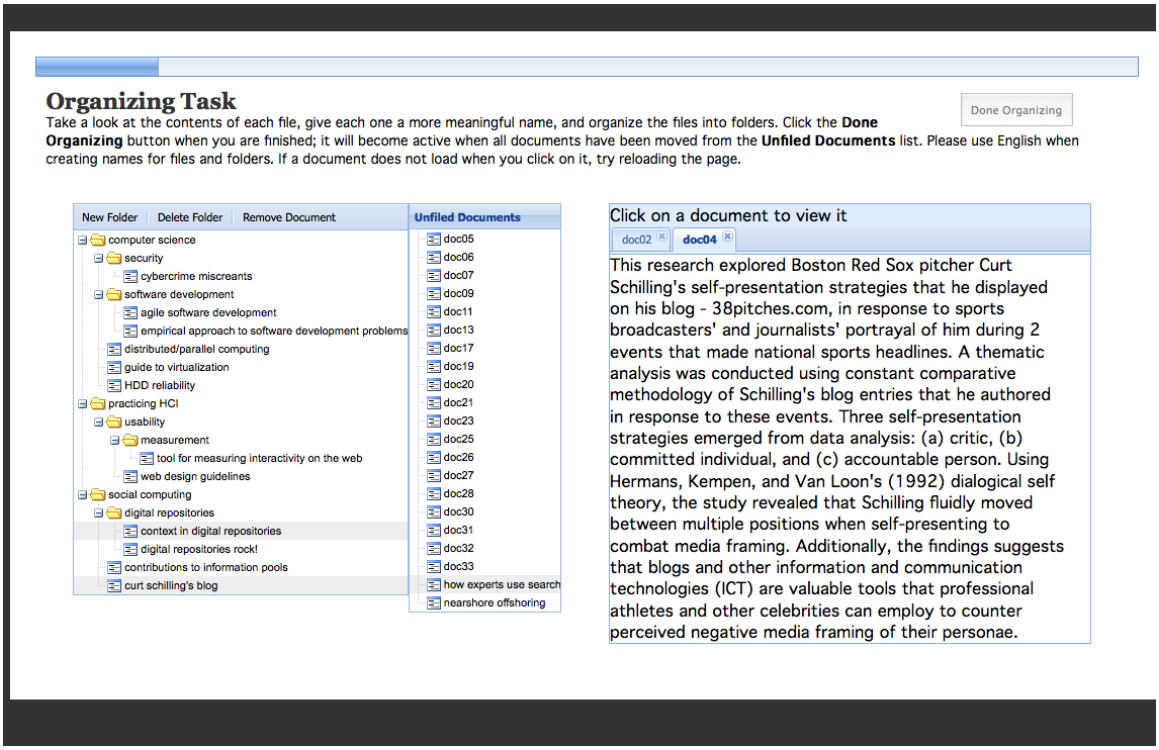


Figure D.2 Labeling and Organizing Interface

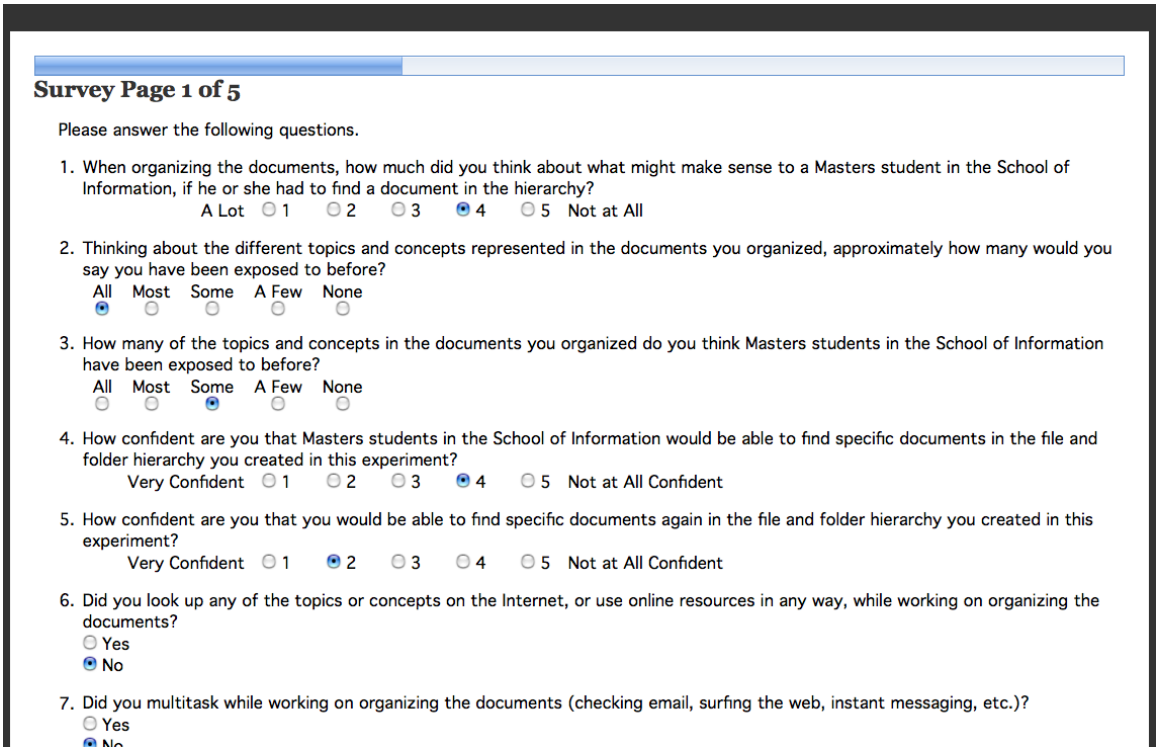


Figure D.3 Example of Questionnaire Interface

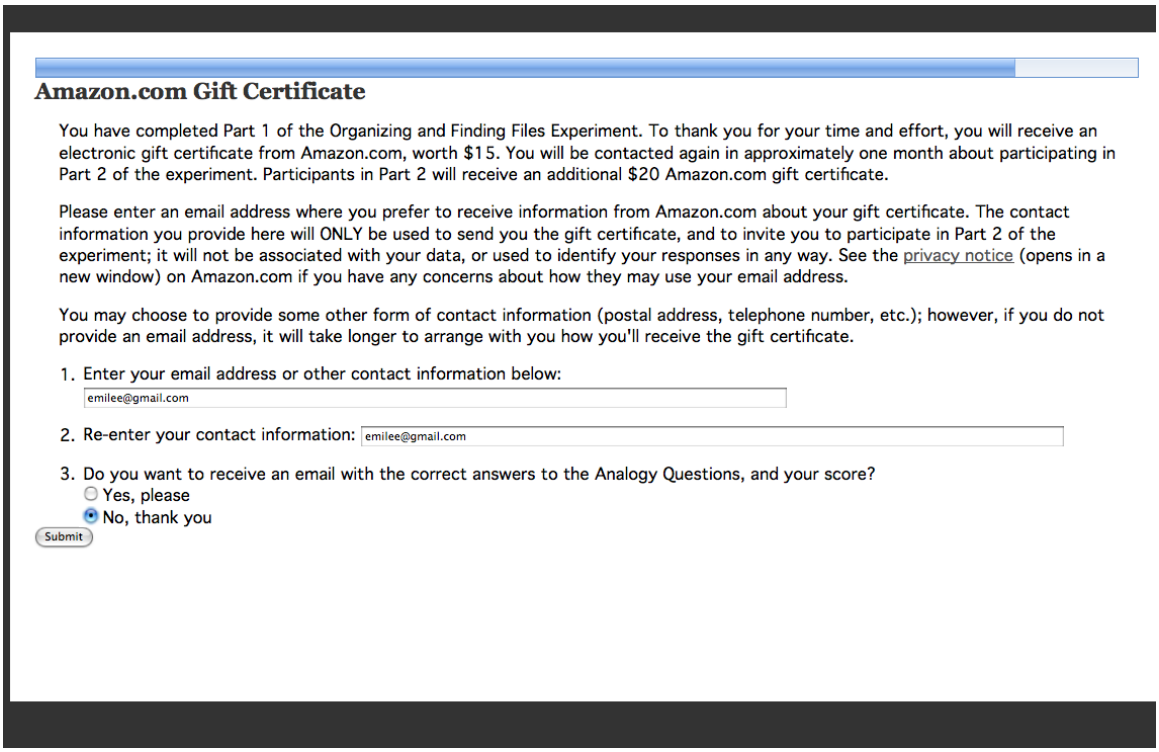


Figure D.4 Thank You Screen

Appendix E

Files for the Organizing and Finding Tasks

The files used in the Organizing and Finding Experiments are short article summaries or excerpts, taken from recent issues of online periodicals and journals. The sources from which the files were selected are:

- Communications of the ACM
- Journal of Computer-Mediated Communication
- Computers in Libraries
- ACM Queue
- Computers in Human Behavior
- ACM interactions

A sample consisting of approximately 50 excerpts were selected by the experimenter such that they all pertain loosely to current issues in Information and/or Computer Science. Some are more relevant to IS graduate students, some to CS graduate students, and some are potentially interesting to both communities. The excerpts were also chosen to minimize the use of specialized vocabulary wherever possible—the topics were intended to be high-level enough that participants would spend their time and effort organizing the documents, not attempting to grasp the specifics in each of the texts. As such, in selecting the excerpts the focus was on feature articles on current "hot topics" that might be interesting for participants to read. Out of the 50 initially selected, 33 were randomly chosen to be used in the experiment.

File 1 (Basili & Zelkowitz, 2008): Computer science has been slow to adopt an empirical paradigm, even as practically all other sciences have done so. Since the time of Aristotle 2,500 years ago, the "natural sciences" (such as physics and biology) have observed nature in order to determine reality. In computer science this rarely happens. Experimentation generally means the ability to build a tool or system – more an existence proof than experiment. While experiments are often organized around the evaluation of algorithms (such as performance and workflow), little has been done that involves humans (such as the development of high-performance codes and test sets based on specified test criteria). But any future advances in the computing sciences require that empiricism takes its place alongside theory formation and tool development.

Here, we explore how to apply an empirical approach toward understanding important problems in software development. Understanding a discipline demands observation, model building, and experimentation. Empirical study is about building models that express our knowledge of the aspects of the domain of greatest interest to us (such as those that cause us the most problems). Learning involves the encapsulation of knowledge, checking that our knowledge is correct, and evolving that knowledge over time. This experimental paradigm is used in many fields, including physics, medicine, and manufacturing. Like other sciences, many disciplines within computer science (such as software engineering, artificial intelligence, and database design) likewise require an empirical paradigm.

Because software development is a human-based activity, experimentation must deal with the study of human activities. Experimentation, in this context, involves evaluating quantitative and qualitative data to understand and improve what the development staff does (such as defining requirements, creating solutions to problems, and programming).

Experimentation requires real-world laboratories. Developers must understand how to build systems better by studying their own environments. Researchers need laboratories in which to observe and manipulate development variables, provide models to predict the cost and quality of the systems, and identify what processes and techniques are most effective in building the system under the given conditions to satisfy a specific set of goals. Research and development have a synergistic relationship that requires a working relationship between industry and academe.

If we take the example of the development of software systems, the developer needs evidence of what does and does not work, as well as when it works. A software organization must be able to answer questions like: What is the right combination of technical and managerial solutions for my problem and environment? What is the right set of processes for my business? How should they be modified? How do we learn from our successes and failures? And how do we demonstrate sustained, measurable improvement? All must be supported by empirical evidence.

File 2 (Caldas, Schroeder, Mesch, & Dutton, 2008): Will the World Wide Web and search engines foster access to more diverse sources of information, or have a centralizing influence through a 'winner-take-all' process? To address this question, we examined how search engines are used to access information about six global issues (climate change, poverty, HIV/AIDS, terrorism, trade reform, and Internet and society). The study used a combination of webmetric analyses and interviews with experts. From interviews we were able to explore how experts on these topics use search engines within their specialist fields. Using webmetric analysis, we were able to compare the results from a number of search engines and show how the top ranked sites are clustered as well as the distribution of their connectivity. Results suggest that the Web tends to reduce the significance of offline hierarchies in accessing information – thereby "democratizing" access to worldwide resources.

It also seems, however, that centers of expertise progressively refine their specializations, gaining a 'winner-take-all' status within a narrower area. Some limitations of the winner-take-all thesis for access to research are discussed.

File 3 (Bullen, 2008): The wonderful Web 2.0 is a famously slippery concept to define. The very ambiguity of the term is Escheresque, self-referential to its ever-changing meaning. As Tim O'Reilly, CEO of O'Reilly Media, described it, "Like many important concepts, Web 2.0 doesn't have a hard boundary, but rather, a gravitational core." As Illinois State Library's information technology coordinator, I have come to realize that embracing this essential Web 2.0 philosophy is a useful tool in unlocking the true potential of digital collections. In fact, the central premise behind this article is that until we embrace Web 2.0 concepts, digital repositories cannot evolve beyond very useful cataloging tools.

All digital collection and information repository tools excel at collecting and disseminating information about the individual items described in their database systems. However, most lack a coherent mechanism for placing the items in context. Images and objects appear one after another in artificial order, devoid of any descriptive connection to each other or to their historical context.

As professionals, we have embraced the concept of preserving materials as digital objects, defined a wide-ranging series of increasingly sophisticated cataloging (metadata) systems, and developed quality assurance and trusted repository programs. In my very humble opinion, however, we have yet to embrace the next logical step, which is to provide context for the images in our collections. Our image repositories are grouped together with as much relevance as posters in a display at a neighborhood head shop; except for an overarching collection theme, our digital objects exist in splendid isolation, beautifully cataloged but bereft of supporting and defining context.

My interpretation of Web 2.0 is that it is a blurring of boundaries. This is the powerful philosophy that will guide the three parts of this article: joining existing tools to put digital images in context, creating tangential or related stories (which I have taken to referring to as the "long tale"), and enabling audience participation to enrich your content.

File 4 (Sanderson, 2008): This research explored Boston Red Sox pitcher Curt Schilling's self-presentation strategies that he displayed on his blog - 38pitches.com, in response to sports broadcasters' and journalists' portrayal of him during 2 events that made national sports headlines. A thematic analysis was conducted using constant comparative methodology of Schilling's blog entries that he authored in response to these events. Three self-presentation strategies emerged from data analysis: (a) critic, (b) committed individual, and (c) accountable person. Using Hermans, Kempen, and Van Loon's (1992) dialogical self theory, the study revealed that Schilling fluidly moved between multiple positions when self-presenting to combat media framing. Additionally, the findings suggests that blogs and other information and communication technologies (ICT) are valuable tools that professional athletes and other celebrities can employ to counter perceived negative media framing of their personae.

File 5 (Richardson, 2008): Separation of concerns is one of the oldest concepts in computer science. The term was coined by Dijkstra in 1974.1 It is important because it simplifies software, making it easier to develop and maintain. Separation of concerns is commonly achieved by decomposing an application into components. There are, however, crosscutting concerns, which span (or cut

across) multiple components. These kinds of concerns cannot be handled by traditional forms of modularization and can make the application more complex and difficult to maintain.

Examples of crosscutting concerns in enterprise Java applications include transaction management, security, persistence, and application assembly. Scattering the code that handles those concerns across multiple components is undesirable, however. This is because doing so would make each component more complex. Also, duplicating code in multiple modules can cause maintenance problems. Consequently, there has been a flurry of activity in recent years to develop frameworks and new forms of modularity that try to untangle these crosscutting concerns from the application's business logic.

In this article we look at the evolution of enterprise Java frameworks that tackle crosscutting concerns. We address how dissatisfaction with the first-generation frameworks, which were based on the EJB (Enterprise JavaBeans) programming model, prompted the development of dramatically better frameworks. These newer-generation frameworks are based on the POJO (Plain Old Java Object) programming model. Even though the focus of this article is Java, it contains many useful lessons for developers of frameworks and applications written in other languages.

File 6 (Chiou & Lee, 2008): This exploratory study is to analyze the communication differences among viewers of US TV program Friends on Internet discussion forum in the US, Japan, and Taiwan. It intends to establish whether exposure to foreign TV could lead to similar communication content in the context of the virtual community between exporting and importing societies. Content analysis was used in this cross-cultural study, with the aim of understanding the ways in which dialogues posted on various discussion forums differed among the United States, Japan, and Taiwan. The results of this exploratory study support the notion that the process of cultural value influence is more complex than cultural imperialism advocates propose. Audiences respond actively rather than passively to foreign TV programs. Prior information structure of the audience is affecting the interpretation of subsequent information.

File 7 (Williams, Yee, & Caplan, 2008): Online games have exploded in popularity, but for many researchers access to players has been difficult. The study reported here is the first to collect a combination of survey and behavioral data with the cooperation of a major virtual world operator. In the current study, 7,000 players of the massively multiplayer online game (MMO) EverQuest 2 were surveyed about their offline characteristics, their motivations and their physical and mental health. These self-report data were then combined with data on participants' actual in-game play behaviors, as collected by the game operator. Most of the results defy common stereotypes in surprising and interesting ways and have implications for communication theory and for future investigations of games.

File 8 (Cheshire & Antin, 2008): A growing number of systems on the Internet create what we call information pools, or collections of online information goods for public, club or private consumption. Examples of information pools include collaborative editing websites (e.g. Wikipedia), peer-to-peer file sharing networks (e.g., Napster), multimedia contribution sites (e.g. YouTube), and amorphous collections of commentary (e.g., blogs). In this study, we specifically focus on information pools that create a public good. Following current theory and research, we argue that extremely low costs of contribution combined with very large networks of distribution facilitate the production of online information pools – despite an abundance of free-riding behavior. This paper presents results from a

series of Internet field experiments that examine the effects of various feedback mechanisms on repeat contributions to an information pool. We demonstrate that the social psychological benefits from gratitude, historical reminders of past behavior, and ranking of one's contributions relative to those of others can significantly increase repeat contributions. In addition, the context in which individuals interact with the system may partially mitigate the positive effect of some types of feedback on contribution behavior.

File 9 (Dubberly, 2008): In the early 20th century, our understanding of physics changed rapidly; now our understanding of biology is undergoing a similar rapid shift. Freeman Dyson wrote: "It is likely that biotechnology will dominate our lives and our economic activities during the second half of the twenty-first century, just as computer technology dominated our lives and our economy during the second half of the twentieth." Recent breakthroughs in biology are largely about information – understanding how organisms encode it, store, reproduce, transmit, and express it – mapping genomes, editing DNA sequences, mapping cell-signaling pathways.

Changes in our understanding of physics, accompanied by rapid industrialization, led to profound cultural shifts: changes in our view of the world and our place in it. In this context, modernism arose. Similarly, recent changes in our understanding of biology are beginning to create new industries and may bring another round of profound cultural shifts: new changes in our view of the world and our place in it. Already we can see the process beginning. Where once we described computers as mechanical minds, increasingly we describe computer networks with more biological terms – bugs, viruses, attacks, communities, social capital, trust, identity.

Over the past 30 years, the growing presence of electronic information technology has changed the context and practice of design. Changes in the production tools that designers use (software tools, computers, networks, digital displays, and printers) have altered the pace of production and the nature of specifications. But production tools have not significantly changed the way designers think about practice. In a sense, graphic designer Paul Rand was correct when he said, "The computer is just another tool, like the pencil," suggesting the computer would not change the fundamental nature of design. But computer-as-production-tool is only half the story; the other half is computer-plus-network-as-media.

Changes in the media that designers use (the Internet and related services) have altered what designers make and how their work is distributed and consumed. New media are changing the way designers think about practice and creating new types of jobs. For many of us, both what we design and how we design are substantially different from a generation ago.

File 10 (Cymru, 2008): Painted in the broadest of strokes, cybercrime essentially is the leveraging of information systems and technology to commit larceny, extortion, identity theft, fraud, and, in some cases, corporate espionage. Who are the miscreants who commit these crimes, and what are their motivations? One might imagine they are not the same individuals committing crimes in the physical world. Bank robbers and scam artists garner a certain public notoriety after only a few occurrences of their crimes, yet cybercriminals largely remain invisible and unheralded. Based on sketchy news accounts and a few public arrests, such as Mafiaboy, accused of paralyzing Amazon, CNN, and other Web sites, the public may infer these miscreants are merely a subculture of teenagers. In this article we provide insight into the root causes of cybercrime, its participants and their motivations, and we identify some of the issues inherent in dealing with this crime wave.

File 11 (Leong, Howard, & Vetere, 2008): Randomness has long beguiled and fascinated human beings. It is widely used as a powerful computational resource, as mathematicians and scientists use it to encrypt, model, and predict. Artists, on the other hand, have recognized randomness's versatility and ability to provoke, seed, and capture our imagination. They exploit the ephemeral qualities of randomness, utilizing them as creative tools to produce innovative artistic output.

For interaction designers, randomness can be used to enrich designed user experiences. Encounters with randomness exploit our natural urge to interpret and our tendency to try to make sense of things when engaging with content in unpredictable and unexpected ways. As design discourse shifts from "beyond the object" into "experience design," the design of digital devices is increasingly motivated by users and their experiences, and an appreciation of the role of randomness can provide designers with a unique perspective as they grapple with the complexities of "users" and "experiences".

File 12 (Poole, 2008): Keeping up with the rapid pace of change can be a daunting task. Just as you finally get your software working with a new technology to meet yesterday's requirements, a newer technology is introduced or a new business trend comes along to upset the apple cart. Whether your new challenge is Web services, SOA (service-oriented architecture), ESB (enterprise service bus), AJAX, Linux, the Sarbanes-Oxley Act, distributed development, outsourcing, or competitive pressure, there is an increasing need for development methodologies that help to shorten the development cycle time, respond to user needs faster, and increase quality all at the same time.

An emerging response to this challenge is a methodology called agile software development, the common theme of which is taking a traditional development process with a single deliverable at the end and splitting it into a series of small iterations, each of which is a microcosm of the full process and each producing working software.

Adopting an agile methodology poses its own set of challenges. It is used mostly by early adopters with small colocated teams, it has little tool support, and though the adoption can be done in phases, getting the full benefits of agile development requires sweeping changes to all phases of the project lifecycle. In addition, agile is actually a large and growing family of methodologies, including extreme programming, Scrum, lean software development, Crystal methodologies, feature-driven development, and many others. Although it is nice to have a wide selection, it does complicate the adoption process. Agile also challenges many fundamental and long-held legacy beliefs about the software development lifecycle. This makes it difficult to achieve the critical mass required to start an agile project.

File 13 (M. K. Chang & Law, 2008): A number of diagnostic scales have been developed in recent years to assess Internet addiction. To better understand the structure, validity, and reliability of such assessment instruments, Young's Internet Addiction Test (IAT) was evaluated using a confirmatory approach.

Data collected through a survey of 410 Hong Kong university undergraduates was subjected to exploratory factor analysis and data from a hold-out sample was analyzed using confirmatory factor analysis in order to assess the psychometric properties and factor structure of the IAT scale. Three dimensions, namely, "Withdrawal and Social Problems", "Time Management and Performance", and "Reality Substitute" were extracted.

These dimensions were then correlated with a number of criterion variables, including academic performance, online activities, gender, and Internet usage. The results show that academic performance was negatively correlated with the Internet addiction scores. The degree of Internet addiction was also found to vary across different types of online activity, with people engaged in cyberrelationships and online gambling having higher Internet addiction scores.

File 14 (McMillan, Hoy, Kim, & McMahan, 2008): Analysis of interactivity in Web sites is an important extension of a long tradition of analyzing content of media messages. But both interactivity and online analysis of content and features offer unique challenges to researchers. This study develops and tests a tool for measuring interactivity in the context of health-related Web sites. The tool was flexible enough to distinguish among multiple types of interactivity and powerful enough to show differences in interactivity based on domain type. Thus, it should have a relatively long life as a multifaceted tool for the tough job of measuring interactivity online.

File 15 (Crosby & Brown, 2008): Are hypervisors the new foundation for system software? A number of important challenges are associated with the deployment and configuration of contemporary computing infrastructure. Given the variety of operating systems and their many versions – including the often-specific configurations required to accommodate the wide range of popular applications – it has become quite a conundrum to establish and manage such systems.

Significantly motivated by these challenges, but also owing to several other important opportunities it offers, virtualization has recently become a principal focus for computer systems software. It enables a single computer to host multiple different operating system stacks, and it decreases server count and reduces overall system complexity. EMC's VMware is the most visible and early entrant in this space, but more recently XenSource, Parallels, and Microsoft have introduced virtualization solutions. Many of the major systems vendors, such as IBM, Sun, and Microsoft, have efforts under way to exploit virtualization. Virtualization appears to be far more than just another ephemeral marketplace trend. It is poised to deliver profound changes to the way that both enterprises and consumers use computer systems.

What problems does virtualization address, and moreover, what will you need to know and/or do differently to take advantage of the innovations that it delivers? In this article we provide an overview of system virtualization, taking a closer look at the Xen hypervisor and its paravirtualization architecture. We then review several challenges in deploying and exploiting computer systems and software applications, and we look at IT infrastructure management today and show how virtualization can help address some of the challenges.

File 16 (Carmel & Abbott, 2008): As the outsourcing and offshoring phenomena matured, the marketplace has sought increased differentiation on the basis of location through a range of 'shoring' and 'sourcing' terms. "Rural-sourcing," "two-shoring," "best-shoring," and at least a dozen other expressions have emerged. Prominent among these is "nearshore," which first appeared in the software/IT field in an article about an entrepreneurial software development venture established in the island of Barbados. Nearshore was presented then as a reaction to the main offshore destination, India, which was viewed as "farshore," a very distant destination, many hours to travel, many time zones away, and a very different culture.

Countries and companies viewing themselves as nearshore claim to offer some of the benefits of offshoring (namely, cost reduction), while mitigating difficulties imposed by distance from the client. Studies on distributed software development have documented that distance introduces difficulties due to issues of communication, control and supervision, coordination, creating social bonds, and building trust. The emergence of nearshoring in an industry that encourages virtual forms of working presents yet more evidence that distance still matters. In this article we explore the subtle ways in which this is viewed.

Traditional offshoring is enabled by technology. The ubiquitous nature of technology has led to an assumption that common interactions such as communication, coordination, and collaboration can be easily resolved over distance by technology and that physical location therefore becomes a non-issue. Such a view is espoused, for example, in the book, The Death of Distance, which claims that "companies will locate any screen-based activity anywhere on earth, wherever they can find the best bargain of skills and productivity". Nearshoring challenges this assumption. Nearshore emphasizes location and proximity as opposed to the prevailing offshoring archetypes of location transparency and irrelevance of distance and time.

Research in related areas, such as the distributed organization of work, global strategy, and economic geography, support the view that despite current globalization trends, location, and distance still do matter. Kiesler and Cummings, investigating geographic distribution of work, assert that proximity is critical to the development of group interaction and social relationships, and that technology alone is often insufficient to re-create the same facilitating environment in distributed teams that is present in co-located settings. Porter, in his landmark article on geographic clustering of related industries such as Silicon Valley, argues that despite the apparent global availability of capital, goods, and information, there is still evidence of competitive advantage based on particularities of location, such as knowledge and relationships. Ghemawat argues that a sophisticated analysis of distance based on several dimensions (such as cultural, geographic, administrative, and economic) is needed in order to better inform the feasibility of making international investments even within a new economic climate that promotes increased global interaction. Ghemawat's dimensions of distance suggest that the impact of distance on global trading relationships can be measured in more than just geographical terms. Each dimension can either enhance or restrict the effectiveness of the relationship.

File 17 (Garfinkel, 2008): A computer used by Al Qaeda ends up in the hands of a Wall Street Journal reporter. A laptop from Iran is discovered that contains details of that country's nuclear weapons program. Photographs and videos are downloaded from terrorist Web sites. As evidenced by these and countless other cases, digital documents and storage devices hold the key to many ongoing military and criminal investigations. The most straightforward approach to using these media and documents is to explore them with ordinary tools – open the word files with Microsoft Word, view the Web pages with Internet Explorer, and so on.

Although this straightforward approach is easy to understand, it can miss a lot. Deleted and invisible files can be made visible using basic forensic tools. Programs called carvers can locate information that isn't even a complete file and turn it into a form that can be readily processed. Detailed examination of e-mail headers and log files can reveal where a computer was used and other computers with which it came into contact. Linguistic tools can discover multiple documents that refer to the same individuals, even though names in the different documents have different spellings and are in different human languages. Data-mining techniques such as cross-drive analysis can reconstruct social networks – automatically determining, for example, if the computer's previous user was in

contact with known terrorists. This sort of advanced analysis is the stuff of DOMEX, the little-known intelligence practice of document and media exploitation.

The U.S. intelligence community defines DOMEX as "the processing, translation, analysis, and dissemination of collected hard-copy documents and electronic media, which are under the U.S. government's physical control and are not publicly available." That definition goes on to exclude "the handling of documents and media during the collection, initial review, and inventory process." DOMEX is not about being a digital librarian; it is about being a digital detective.

Although very little has been disclosed about the government's DOMEX activities, in recent years academic researchers – particularly those concerned with electronic privacy – have learned a great deal about the general process of electronic document and media exploitation. My interest in DOMEX started while studying data left on hard drives and memory sticks after files had been deleted or the media had been "formatted." I built a system to automatically copy the data off the hard drives, store it on a server, and search for confidential information. In the process I built a rudimentary DOMEX system. Other recent academic research in the fields of computer forensics, data recovery, machine translation, and data mining is also directly applicable to DOMEX.

This article introduces electronic document and media exploitation from that academic perspective. It presents a model for performing this kind of exploitation and discusses some of the relevant academic research. Properly done, DOMEX goes far beyond recovering documents from hard drives and storing them in searchable archives. Understanding this engineering problem gives insight that will be useful for designing any system that works with large amounts of unstructured, heterogeneous data.

File 18 (Tao, 2008): Web design guidelines are adopted by many usability evaluation methods as one of the criteria for success, while usability is proven to significantly impact Website performance. Since Web design guidelines cover a broad range of system and interface design solutions, knowledge of them can be considered as a prominent indicator of Web design skills for information systems (IS) professionals. This study empirically assessed how much IS professionals know and apply Web design guidelines via a survey to 500 randomly selected companies from Taiwan's Fortune 2000 corporations. As expected, the knowledge-application gaps of IS professionals were statistically significant in all Web design guideline categories. *M*while, certain guideline categories were proven to be more difficult to acquire or apply than others. Finally, degree, gender, experience, training hours, and courses taken were also proven to be determining factors for Web design guideline skills. Implications for developing Web design guideline skills are also discussed.

File 19 (Henning, 2008): After more than 25 years as a software engineer, I still find myself underestimating the time it will take to complete a particular programming task. Sometimes, the resulting schedule slip is caused by my own shortcomings: as I dig into a problem, I simply discover that it is a lot harder than I initially thought, so the problem takes longer to solve – such is life as a programmer. Just as often I know exactly what I want to achieve and how to achieve it, but it still takes far longer than anticipated. When that happens, it is usually because I am struggling with an API that seems to do its level best to throw rocks in my path and make my life difficult. What I find telling is that, after 25 years of progress in software engineering, this still happens. Worse, recent APIs implemented in modern programming languages make the same mistakes as their two-decade-old counterparts written in C. There seems to be something elusive about API design that, despite many years of progress, we have yet to master.

File 20 (Garrett & Danziger, 2008): Many contemporary analyses of personal Internet use during work explain the behavior in terms of workplace disaffection. However, evidence for this interpretation is mixed. This article posits that an approach emphasizing the expected outcomes of Internet use more effectively explains the behavior. The 2 approaches are tested using survey data collected from more than 1,000 U.S.-based computer-using workers. About 4/5 of those workers do engage in personal Internet use during work. Regression analyses show that workplace disaffection factors, such as stress and dissatisfaction, have no significant influence on the extent of web surfing or personal e-mail use during work. In contrast, factors which shape the expected outcomes of personal Internet use during work, such as a generalized positive perception of the utility of the Internet, routinized use of computers, job commitment, and organizational restrictions on computer use, are very significant predictors of such behavior. These results suggest that employees use the Internet for personal purposes at work for many of the same reasons that they use it elsewhere. Implications of these findings are explored.

File 21 (Workman, Bommer, & Straub, 2008): Organizations and individuals are increasingly impacted by misuses of information that result from security lapses. Most of the cumulative research on information security has investigated the technical side of this critical issue, but securing organizational systems has its grounding in personal behavior. The fact remains that even with implementing mandatory controls, the application of computing defenses has not kept pace with abusers' attempts to undermine them. Studies of information security contravention behaviors have focused on some aspects of security lapses and have provided some behavioral recommendations such as punishment of offenders or ethics training. While this research has provided some insight on information security contravention, they leave incomplete our understanding of the omission of information security measures among people who know how to protect their systems but fail to do so. Yet carelessness with information and failure to take available precautions contributes to significant civil losses and even to crimes. Explanatory theory to guide research that might help to answer important questions about how to treat this omission problem lacks empirical testing. This empirical study uses protection motivation theory to articulate and test a threat control model to validate assumptions and better understand the "knowing-doing" gap, so that more effective interventions can be developed.

File 22 (Papadopoulos, Bruno, & Katz, 2008): In the early '90s, the Berkeley NOW (Network of Workstations) Project under David Culler posited that groups of less capable machines (running SunOS) could be used to solve scientific and other computing problems at a fraction of the cost of larger computers. In 1994, Donald Becker and Thomas Sterling worked to drive the costs even lower by adopting the then-fledgling Linux operating system to build Beowulf clusters at NASA's Goddard Space Flight Center. By tying desktop machines together with open source tools such as PVM (Parallel Virtual Machine), MPI (Message Passing Interface), and PBS (Portable Batch System), early clusters – which were often PC towers stacked on metal shelves with a nest of wires interconnecting them – fundamentally altered the balance of scientific computing. Before these first clusters appeared, distributed/parallel computing was prevalent at only a few computing centers, national laboratories, and a very few university departments. Since the introduction of clusters, distributed computing is now, literally, everywhere.

There were, however, ugly realities about clusters. The lack of tools meant that building 16 or 32 machines to work closely together was a heroic systems effort. Open source software was (and often still is) poorly documented and lacked critical functionality that more mature commercial products offered on the "big machines." It often took months to get a cluster up and running and took highly

trained experts to get it into that condition. It took even longer for applications to run reasonably well on these cheaper machines, if at all.

Nonetheless, the potential of building scalable and cheap computing was too great to be ignored, and the community as a whole grew more sophisticated until clusters became the dominant architecture in high-performance computing. Midsize clusters are now about 100 machines in strength, big clusters consist of 1,000 machines, and the biggest supercomputers are even larger cluster machines. For HPC (high-performance computing), clusters have arrived. Most of these are either Linux-based or a commercial Unix derivative, with most of the top 500 machines running a Linux derivative. There is a new trend toward better hardware integration in terms of blades. This helps eliminate significant wiring clutter.

The past 12 years of clusters have honed community experience – many can turn out "MPI boxes" (homogeneous hardware that enables message-passing parallel applications), and there are several software tools that understand clusters and allow non-experts to go from bare metal (e.g., that cluster SKU from your favorite computing hardware company) to functioning cluster made up of hundreds of individual nodes (computers) in a few hours. At the National Center for Supercomputing Applications, the Tungsten2 Cluster (512 nodes) went from purchase order placement to full production in less than a month and was one of the 50 fastest supercomputers in the world in June 2005.

It seems that the problems with clusters have been solved, but their wild success means that everyone wants to do more with them. While retaining roots born in HPC, clusters of Web servers, tiled display walls, database servers, and file servers are becoming commonplace. Nearly every entity in the modern machine room is essentially a clustered architecture. Building a specialized MPI box (classic Beowulf cluster) is a small subset of what is needed to support the needs of computational researchers.

File 23 (Jagatic, Johnson, Jakobsson, & Menczer, 2008): Phishing is a form of deception in which an attacker attempts to fraudulently acquire sensitive information from a victim by impersonating a trustworthy entity. Phishing attacks typically employ generic "lures." For instance, a phisher misrepresenting himself as a large banking corporation or popular online auction site will have a reasonable yield, despite knowing little to nothing about the recipient. In a study by Gartner Group, about 19% of all those surveyed reported having clicked on a link in a phishing email message, and 3% admitted to giving up financial or personal information. The research project described here was designed to provide us with a baseline success rate for individual phishing attacks, and was, when it was performed in 2005, the first study to achieve this goal.

It is worth noting that phishers are getting smarter. Following trends in other online crimes, it is inevitable that future generations of phishing attacks will incorporate greater elements of context to become more effective and thus more dangerous for society. For instance, suppose a phisher were able to induce an interruption of service to a frequently used resource, for example, to cause a victim's password to be locked by generating excessive authentication failures. The phisher could notify the victim of a "security threat." Such a message may be welcomed or expected by the victim, who would then be easily induced into disclosing personal information.

In other forms of so-called spear phishing or context aware phishing, an attacker would gain the trust of victims by obtaining information about their bidding history or shopping preferences (freely available from eBay), their banking institutions (discoverable through their Web browser history, see browser-recon.info), or their mothers' maiden names (which can be inferred from data required by law to be public). Avi Rubin, a computer science professor at Johns Hopkins University, designed a

class project for his graduate course, "Security and Privacy in Computing," to demonstrate how a database can be built to facilitate identity theft. The project focused on residents of Baltimore using data obtained from public databases, Web sites, public records, and physical world information that can be captured on the computer.

Given that phishing attacks take advantage of both technical and social vulnerabilities, there are a large number of different attacks. Here, we discuss how phishing attacks can be honed by means of publicly available personal information from social networks. The idea of using people's social contacts to increase the power of an attack is analogous to the way in which the "ILOVEYOU" virus used email address books to propagate itself. The question we ask here is: How easily and effectively can a phisher exploit social network data found on the Internet to increase the yield of a phishing attack? The answer, as it turns out, is very easily and very effectively. Our study suggests that Internet users may be over four times as likely to become victims if they are solicited by someone appearing to be a known acquaintance.

File 24 (Huwe, 2008): The pace of online "content creation" has never been faster. The explosive growth in blogs and their growing influence, not to mention the money they create for entrepreneurs, is just the most obvious case in point. We now take for granted the prompt creation of high-quality content from myriad think tanks, advocacy groups, and even mainline newspapers such as The New York Times. Massive ebook repositories such as Google Scholar finish off the picture, and the media see them as inevitable. But there's one sector of digital content creation that receives less mainstream attention, and you may not even know much about it unless you work in academia: the digital repository. I see a great deal of long-term value in this area of digital library development. And lately, my experience with it has been very direct, very engrossing, and, not least, an intellectual challenge.

Unlike blogs with accretions of documents and commentary or websites with masses of ephemeral pages that come and go, digital repositories have a more plodding pace of development. They conform to demanding standards for metadata and information architecture–and those standards are still evolving. They often operate on open source platforms and are attached to research universities or nonprofit outfits. But despite the pedestrian, even dull, aspects of repository development, announcing new repositories can cause a big splash. I'll talk more about that later.

Repository development also has an unexpected benefit: It reinvigorates the best in our long-term professional values and makes them understandable for contemporary society. Good repository projects take root at the local level and follow the interests of researchers or scholars who care a great deal about very narrow slices of the information universe. From this specialized environment, a digital repository can spring into being as a fully operational digital tool, which has already enjoyed years of careful organization. My view is that we are often storing treasures in musty corners; some of you may agree. Never mind how obscure the source is. For example, when was the last time you, the reader, studied a collective bargaining agreement? It's unlikely you ever have–but in my case, I can name a dozen local scholars, and more around the globe, who do exactly that. Art historians, immigration researchers, anthropologists, and legal scholars can tell similar stories about the hidden treasures in nearly forgotten sources.

File 25 (Goeke & Faley, 2008): To better harness the power of their own data, many firms have invested in data warehousing technology. A data warehouse enables the collection and storage of vast amounts of data that is then extracted and analyzed by end users. Sophisticated software allows data mining, and gives analysts and managers unprecedented visibility into business operations.

For instance, business intelligence software can use the data warehouse to track results as well as better understand the factors that led to them. Moreover, sales can be enhanced via customer relationship management, which uses the data warehouse to extract and explain complex buyer behaviors. It is no surprise that with its potential to generate extraordinary gains in productivity and sales, the market for data warehousing software and hardware was nearly $200 billion by 2004. The market for data warehousing technology will continue to expand because businesses highly regard its decision-support functionality and because it has been adapted to better meet the needs of a growing number of small- and medium-sized businesses.

Potential benefits notwithstanding, data warehouses are still large, expensive, and risky undertakings. For example, the median installation cost of a data warehouse is $1.5 million and can exceed $50 million, which does not include annual operating costs. In addition, it has been noted that 40% of all data warehousing implementations fail. Even if the implementation succeeds, end-user success with the data warehouse can still be problematic.

Indeed, a troubling pattern in the growing literature about data warehousing is that end users experience substantial difficulty with their firms' data warehouses. This seems paradoxical, as data warehouses are engineered to be user-driven, allowing end users to be in control of their data. It is this paradox – that data warehouses are used even though end users find them difficult – we explore in this article.

File 26 (Popovich, Gullekson, Morris, & Morse, 2008): The importance and use of computers has increased dramatically over the last two decades. The Attitudes Towards Computer Usage Scale (ATCUS) was developed in 1986 [Popovich, P. M., Hyde, K. R., Zakrajsek, T., & Blumer, C., (1987). The development of the attitudes toward computer usage scale. Educational and Psychological Measurement, 47, 261-269.] and used in a variety of settings over the years. In order to examine how computer attitudes have changed from 1986 to 2005, the ATCUS was given to 254 male and female current undergraduate students. When comparing the 1986 with 2005 results, the amount of time spent using a computer was still positively related to computer attitudes; however, the number of college computer courses was not. There is no longer a significant relationship among any of the factors with college computer courses. Males and females no longer significantly differ in their attitudes toward computers, number of college computer courses, amount of time spent using computers, or degree of self-reported computer anxiety. Implications are discussed.

File 27 (Steinman, 2008): Communications systems based on the ISP (Session Initiation Protocol) standard have come a long way over the past several years. ISP is now largely complete and covers even advanced telephony and multimedia features and feature interactions. Interoperability between solutions from different vendors is repeatedly demonstrated at events such as the ISPit (interoperability test) meetings organized by the ISP Forum, and several manufacturers have proven that proprietary extensions to the standard are no longer driven by technical needs but rather by commercial considerations.

Even in light of all this excellent news, most implementations still fall short in one key area: native ISP call control and ISP-based feature interaction required for multivendor interoperability. ISP first unfolds its full potential if it is used for more than just transport "channels" that interconnect otherwise proprietary IP PBX implementations. The simple fact that ISP "goes in" and "comes out" of a PBX system does not mean that this system has much to do with ISP at all. Native ISP call control and ISP transport "channels" are two very different and oft-confused architectural approaches

to building ISP communications systems.

Thanks to many enterprise users who increasingly insist on standards-based, open, and therefore interoper-able systems, the industry is embracing a new model. Realtime communications, including telephony, is starting to look like yet another IT application – an application that runs on standard hardware, uses standard operating systems and middleware, and follows PC-like economics. It is an application designed as an open system that accommodates a wide variety of endpoints from many different vendors and integrates into an existing IT infrastructure with Web services, corporate directories, and IT best practices.

File 28 (Callow, Beardow, & Brittain, 2008): One thing that becomes immediately apparent when creating and distributing mobile 3D games is that there are fundamental differences between the cellphone market and the more traditional games markets, such as consoles and handheld gaming devices. The most striking of these are the number of delivery platforms; the severe constraints of the devices, including small screens whose orientation can be changed; limited input controls; the need to deal with other tasks; the nonphysical delivery mechanism; and the variations in handset performance and input capability.

Outside of the mobile market, developers had to target only two or three devices and deliver games on high-capacity media; in the mobile market they have to consider tens of devices per operator and package their games to fit in a compact download. The number and types of devices are also constantly changing; in a 12-month development cycle many new handsets can emerge. In addition, console development focus has been about length of play and numbers of levels, rather than the short bursts of intense activity that typify today's mobile games. Furthermore, development time and budget are usually limited by low retail prices in the $5-10 range.

File 29 (Elerath, 2008): HDDs (hard-disk drives) are like the bread in a peanut butter and jelly sandwich – sort of an unexciting piece of hardware necessary to hold the "software." They are simply a means to an end. HDD reliability, however, has always been a significant weak link, perhaps the weak link, in data storage. In the late 1980s people recognized that HDD reliability was inadequate for large data storage systems so redundancy was added at the system level with some brilliant software algorithms, and RAID (redundant array of inexpensive disks) became a reality. RAID moved the reliability requirements from the HDD itself to the system of data disks. Commercial implementations of RAID range from n+1 configurations (mirroring) to the more common RAID-4 and RAID-5, and recently to RAID-6, the n+2 configuration that increases storage system reliability using two redundant disks (dual parity). Additionally, reliability at the RAID group level has been favorably enhanced because HDD reliability has been improving as well.

Seagate and Hitachi both have announced plans to ship one-terabyte HDDs by the time this article appears.1 With higher areal densities, lower fly-heights (the distance between the head and the disk media), and perpendicular magnetic recording technology, can HDD reliability continue to improve? The new technology required to achieve these capacities is not without concern. Are the failure mechanisms or the probability of failure any different from predecessors? Not only are there new issues to address stemming from the new technologies, but also failure mechanisms and modes vary by manufacturer, capacity, interface, and production lot.

How will these new failure modes affect system designs? Understanding failure causes and modes for HDDs using technology of today and the near future will highlight the need for design alternatives

and tradeoffs that are critical to future storage systems. Software developers and RAID architects can not only better understand the effects of their decisions, but also know which HDD failures are outside their control and which they can manage, albeit with possible adverse performance or availability consequences. Based on technology and design, where must they place the efforts for resiliency?

This article identifies significant HDD failure modes and mechanisms, their effects and causes, and relates them to system operation. Many failure mechanisms for new HDDs remain unchanged from the past, but the insidious undiscovered data corruptions (latent defects) that have plagued all HDD designs to one degree or another will continue to worsen in the near future as areal densities increase.

File 30 (Barr & Cabrera, 2008): In the 50 years since John McCarthy coined the term artificial intelligence, much progress has been made toward identifying, understanding, and automating many classes of symbolic and computational problems that were once the exclusive domain of human intelligence. Much work remains in the field because humans still significantly outperform the most powerful computers at completing such simple tasks as identifying objects in photographs – something children can do even before they learn to speak.

Software developers with innovative ideas for businesses and technologies are constrained by the limits of artificial intelligence. In today's business landscape where companies are more cost-conscious than ever, projects that require a vast network of humans are scrutinized with a fine-tooth comb and often scrapped because the cost of establishing and managing a network of skilled people to do the work outweighs the value of completing it. If software developers could programmatically access and incorporate human intelligence into their applications, a whole new class of innovative businesses and applications would be possible. This is the goal of Amazon Mechanical Turk:1 to give software developers and businesses the power to use human intelligence as a core component of their applications and businesses. With Amazon Mechanical Turk, people are freer to innovate because they can now imbue software with real human intelligence.

In 1769, Wolfgang von Kempelen built an automaton that defeated many human opponents at chess. Known as "The Turk," the wooden mannequin toured the United States and Europe for many years, defeating such famous challengers as Benjamin Franklin, Napolean Bonaparte, and Edgar Allen Poe.2 The secret to the automaton was, of course, a human chess master hidden inside. Like its namesake, Amazon's Mechanical Turk presents a mechanical front to conceal, or abstract, the human processing power and intelligence hidden inside. Developers can use the Amazon Mechanical Turk Web services API to submit tasks to the Amazon Mechanical Turk Web site, approve completed tasks, and incorporate the answers into their software applications. To the application, the transaction looks very much like any remote procedure call: The application sends the request, and the service returns the results. In reality, a network of humans fuels this "artificial artificial intelligence" by coming to the Web site, searching for and completing tasks, and receiving payment for their work. This allows software developers to easily and economically build programs that tap into a worldwide, massively parallel, Internet-scale human workforce on an incremental, as-needed basis.

File 31 (Anthias & Sankar, 2008): Companies have always been challenged with integrating systems across organizational boundaries. With the advent of Internet-native systems, this integration has become essential for modern organizations, but it has also become more and more complex, especially as next-generation business systems depend on agile, flexible, interoperable, reliable, and secure cross-enterprise systems.

This article describes the various demanding scenarios in the cross-enterprise domain and offers perspectives in addressing these challenges. We look at the trajectory and locus of cross-enterprise systems, the many ways in which the various complexities are addressed now, and how they can be simplified in the future. It is in this context that the network emerges as one of the alternatives, acting as an intermediary for cross-enterprise integration of federated business services, with an application orientation.

File 32 (Geer, 2008): Will security threats bring an end to general-purpose computing? Inflection points come at you without warning and quickly recede out of reach. We may be nearing one now. If so, we are now about to play for keeps, and "we" doesn't mean just us security geeks. If anything, it's because we security geeks have not worked the necessary miracles already that an inflection point seems to be approaching at high velocity.

Many of us believe and many more of us say that complexity and security are antipodal. This complexity vs. security dichotomy is real but not exact; yet it is to some degree measurable, and news from that front is not good. The software industry sells a product that does not naturally wear out and that retains complete fidelity when copied – two characteristics, among others, that separate the digital world from the physical world. To continue to make money from existing customers, a software supplier must sell upgrades, maintenance, or both. Maintenance sells best when a product is unstable or hard to use – the very need for maintenance is an admission of complexity. New features, if they are to compel otherwise happy users to effectively repurchase a product they already have, tend to be at least linear (10 new features) if not geometric (10% new features). Absent perfection, each new feature comes with new failure modes, and features can sometimes interact; therefore, the potential number of failure modes quite naturally can grow faster than the feature count.

File 33 (West, 2008): "... [the system] must be easy to use and must neither require stress of mind nor the knowledge of a long series of rules..." –Auguste Kerckhoffs on the design of cryptographic systems (La cryptographie militaire, 1883)

The importance of the user in the success of security mechanisms has been recognized since Auguste Kerckhoffs published his treatise on military cryptography, La cryptographie militaire, over a century ago. In the last decade, there has been tremendous increase in awareness and research in user interaction with security mechanisms.

Risk and uncertainty are extremely difficult concepts for people to evaluate. For designers of security systems, it is important to understand how users evaluate and make decisions regarding security. The most elegant and intuitively designed interface does not improve security if users ignore warnings, choose poor settings, or unintentionally subvert corporate policies. The user problem in security systems is not just about user interfaces or system interaction. Fundamentally, it is about how people think of risk that guides their behavior. There are basic principles of human behavior that govern how users think about security in everyday situations and shed light on why they undermine security by accident.

Appendix F

Questionnaire and Results

This Appendix includes a series of questions participants answered as part of the Organizing and Finding experiments, along with descriptive statistics for each question. If any statistical differences between conditions were detected, those results are included as well.

Participants in the Organizing Phase, by Community Membership and Target Audience:

	CS	IS
Different	11	11
None	10	10
Same	9	10
Self	11	12

Participants in the Finding Phase, by Community Membership:

CS	IS
24	24

Your Gender:

	CS	IS
Female	3	37
Male	38	6

Is English your first language?

	CS	IS
No	17	5
Yes	24	38

How long have you been a student at the University of Michigan? *2-6 months (5), 7-12 months (4), 1-2 years (3), 3-4 years (2), 5 or more years (1)*
CS vs. IS: Kruskal-Wallis W = 25.1704, df = 1, $p < 0.0001$

	M	*SD*
CS	2.10	1.22
IS	3.81	1.40

Thinking about all the different projects or assignments you are currently working on, about how many of them involve collaboration with others? *All (1), Most (2), Some (3), A few (4), none (5), don't know (6)*

	M	*SD*
CS	2.63	0.97
IS	2.47	0.83

Overall, how often do you use the internet? *Hourly or more often (7), Several times a day (6), About once a day (5), 3-5 days a week (4), 1-2 days a week (3), Every few weeks (2), Less often (1)*

	M	*SD*
CS	6.68	0.47
IS	6.49	0.55

Below is a short list of activities people sometimes do online. Please indicate whether you have ever done each one, or not. Have you ever...? *Yes (1), No (2), Don't know (3)*; Where the counts in the table below do not add up to 84 (the total number of participants in the Organizing Phase), the remaining answers were "Don't know".

	Question	Comp. Sci.		Info. Sci.	
		No	Yes	No	Yes
1	Created or worked on your own online journal or blog	18	22	12	31
2	Created or worked on your own webpage	8	33	12	30
3	Read the online journals or blogs of others	4	37	0	43
4	Created or worked on web pages for others, including friends, groups you belong to, or for school projects	8	33	6	37
5	Shared something online that you created yourself, such as your own artwork, photos, stories or videos	6	35	4	38
6	Collaborated on a document or spreadsheet online using Google Docs or something similar	10	30	2	41
7	Used a shared calendar like Microsoft Exchange or Google Calendar or something similar	9	32	1	42
8	Posted a message to an online forum, newsgroup, or email list	4	37	4	37
9	Shared files with others online using a system like CTools, Subversion or CVS, or shared network folders	2	39	1	42

The documents were interesting to read. *5 point likert scale: Strongly Disagree (1) — Strongly Agree (5)*

	M	*SD*
CS	2.83	0.97
IS	3.19	1.01

When organizing the documents, how much did you think about what might make sense to (*you -OR- the target audience*), if he or she had to find a document in the hierarchy? *5 point likert scale: A Lot(1) — Not at All (5)*

SAME **Mean=2.63** **St.Dev.=1.16**	NS	NS	NS
Kruskal-Wallis W=0.23, df=1, p=0.63	**DIFFERENT** **Mean=2.77** **St.Dev.=1.19**	significant difference: Different vs. Self	significant difference: Different vs. None
Kruskal-Wallis W=2.36, df=1, p=0.12	Kruskal-Wallis W=4.02, df=1, p=0.04	**SELF** **Mean=2.09** **St.Dev.=0.90**	NS
Kruskal-Wallis W=2.95, df=1, p=0.09	Kruskal-Wallis W=4.69, df=1, p=0.03	Kruskal-Wallis W=0.075, df=1, p=0.78	**NONE** **Mean=2.00** **St.Dev.=0.86**

Figure F.1 Audience Importance Question: Pairwise Comparisons

Thinking about the different topics and concepts represented in the documents you organized, approximately how many would you say you have been exposed to before? *All (1), Most (2), Some (3), A Few (4), None (5)*

CS vs. IS: Kruskal-Wallis W = 9.1916, df = 1, p-value = 0.002

	M	*SD*
CS	2.24	0.80
IS	2.81	0.79

How many of the topics and concepts in the documents you organized do you think (target audience) has been exposed to before? *All (1), Most (2), Some (3), A Few (4), None (5)*

SAME **Mean=2.42** **St.Dev.=0.51**	NS	NS	NS
Kruskal-Wallis W=2.50, df=1, p=0.1139	**DIFFERENT** **Mean=2.09** **St.Dev.=0.81**	significant difference: Different vs. Self	NS
Kruskal-Wallis W=2.88, df=1, p=0.09	Kruskal-Wallis W=7.83, df=1, p=0.0051	**SELF** **Mean=2.82** **St.Dev.=0.78**	significant difference: Self vs. None
Kruskal-Wallis W=2.64, df=1, p=0.10	Kruskal-Wallis W=0.17, df=1, p=0.68	Kruskal-Wallis W=8.75, df=1, p=0.003	**NONE** **Mean=2.15** **St.Dev.=0.48**

Figure F.2 Topic Familiarity Question: Target Audience Pairwise Comparisons

Did you look up any of the topics or concepts on the Internet, or use online resources in any way, while working on organizing the documents? *Yes or No*

	CS (organizing)	IS (organizing)	CS (finding)	IS (finding)
No	36	40	23	24
Yes	4	3	1	0

Instructions: Analogy questions test the ability to recognize the relationship that exists between the words in a word pair, and to recognize when two word pairs display parallel relationships. In each of the following questions, a word pair in capital letters is followed by five lettered word pairs. Choose the lettered pair that expresses a relationship most similar to the relationship expressed in the capitalized pair. Please do NOT look up the correct answers on the Internet before completing the questions. Most people find these kinds of questions to be quite difficult, so don't worry if the correct answers aren't obvious to you. If you're curious about how you did, you may request that your score and the correct answers be emailed to you after the experiment.

	M	*SD*
CS	2.73	1.07
IS	3.05	1.09

EXPEL : PUPIL ::
inter : prisoner
question : inquisitor
deport : alien (correct answer)
instigate : firebrand
judge : defendant

PRESERVE : MORATORIUM ::
tyrannize : revolt
shade : tree (correct answer)
solve : problem
accumulate : collection
cover : eclipse

PERJURY : OATH ::
plagiarism : authority
embezzlement : trust (correct answer)
disrespect : age
testimony : court
jury : vow

ANECDOTE : STORY ::
ballad : song
novel : chapter
limerick : poem (correct answer)
prose : poetry
overture : opera

COLOR : SPECTRUM ::
tone : scale (correct answer)
sound : waves
verse : poem
dimension : space
cell : organism

Appendix G

Hierarchy Measures and Results

This Appendix includes descriptive statistics for several different hierarchy measures. In addition, if any statistical differences between conditions were detected, those results are included as well. The homogeneity of variance and normality assumptions were violated for most of these measures, except where I state to the contrary below. Because I use nonparametric statistical tests, there are no tests for interaction effects.[1]

There are two categories of measures reported here. The first are measures of individual hierarchies; for example, folder.count. One must only count the number of folders in a particular hierarchy to calculate a score on this measure for that hierarchy.

The other category of measures are what I call *comparison measures*, because they are measures of the similarity or difference between two hierarchies. For example, interuser agreement (G. Furnas et al., 1983) is a measure of the overlap in the words chosen by two participants to represent a particular file in the experiment; this measure cannot be calculated for a single hierarchy in isolation. The measures are described in this appendix, but the results are discussed in Chapter 6.

Number of participants:

	CSE	MSI
Different	11	11
None	10	10
Same	9	10
Self	11	12

[1] I have not been able to discover any reliable nonparametric tests for interaction effects.

G.1 Topology Measures

The measures below all capture some aspect of the structure of a hierarchy, ignoring the specific vocabulary choices for the file and folder labels, and any semantic meaning or subjective rules behind what makes files "go together" into folders.

folder.count: the number of folders in the hierarchy

	M	*SD*
Different	9.09	2.89
None	8.85	2.58
Same	10.11	4.70
Self	9.52	4.52

doc.adj: a score for each hierarchy based on a file-by-file matrix with 33^2 cells. Each cell in the matrix contains a '1' or '0' depending on whether two files are grouped in the same folder in that particular hierarchy. Lower numbers mean fewer files grouped together with other files; essentially, more sprawling hierarchies. Higher numbers mean fewer folders with more files

	M	*SD*
Different	86.27	33.48
None	92.90	47.07
Same	101.63	111.32
Self	96.96	64.62

avg.path: average pairwise document distance, based on the number of steps from every file to every other file in a given hierarchy. This is an indication of structural complexity; a higher average means in general, files are further apart.

	M	*SD*
Different	3.88	0.49
None	3.83	0.51
Same	4.00	0.67
Self	3.86	0.50

folder.mean.out.degree: this measure represents the mean number of children for each folder in a hierarchy, including both files and folders. This measure is an indication of the breadth of a hierarchy.

	M	*SD*
Different	4.71	1.19
None	4.96	1.26
Same	5.92	6.66
Self	5.55	3.10

file.mean.depth: the average depth of files in a hierarchy.

	M	*SD*
Different	2.36	0.64
None	2.48	0.53
Same	2.34	0.47
Self	2.46	0.52

file.adjacency: a *comparison measure*. Create a file-by-file matrix for each hierarchy created in the organizing phase; cells contain '1' if the pair of files are in the same folder, '0' if they are not. Then, the pair of matrices that are being compared are "stacked up", and corresponding cells are added together. In the matrix that results, count the number of cells with '2's and divide by the number of non-zero cells.

This overall "score" for each hierarchy comparison represents the proportion of documents both users put together in the same folder. Two identical matrices, i.e. a hierarchy compared with itself, has a total score of 1.

G.2 Vocabulary Measures

The measures below focus on the specific vocabulary choices participants made when organizing.

mean.rank: to calculate this measure, first make a list of all the different words used in file and folder labels, by all participants in the experiment. After stemming and stop word removal, order them from most to least frequent in the corpus, with rank 1 representing the most common word. Then, for each participant, compute the average rank for all the words they used. Higher average scores mean more unique words (Krauss et al., 1968).

The community membership x target audience ANOVA showed a main effect of target

	M	*SD*
Different	43.22	6.90
None	38.95	4.74
Same	39.08	4.97
Self	42.93	6.51

audience on the mean.rank measure.

	Df	Sum Sq	*M* Sq	F value	Pr(>F)
community	1	18.98	18.98	0.56	0.4567
audience	3	345.25	115.08	3.39	**0.0221**
community $*$ audience	3	209.53	69.84	2.06	0.1127
Residuals	76	2577.47	33.91		

unique.words: the total number of unique words used by a participant in their file and folder labels, relative to all the other participants in the experiment (Krauss et al., 1968).

	M	*SD*
Different	4.00	3.46
None	2.90	2.34
Same	3.47	4.77
Self	4.65	3.97

file.mean.label.len.chars: total number of characters in a filename.

	M	*SD*
Different	26.48	21.54
None	23.43	8.14
Same	24.71	8.34
Self	29.76	9.94

filename.ratio: the ratio of the number of unique words used by one participant in their hierarchy file labels, to the total number of words used (Krauss et al., 1968).

	M	*SD*
Different	0.88	0.04
None	0.86	0.05
Same	0.86	0.04
Self	0.86	0.04

folder.ratio: same as the previous measure, but using words in folder labels rather than file labels.

	M	*SD*
Different	0.93	0.10
None	0.90	0.09
Same	0.93	0.08
Self	0.92	0.07

label.file.similarity: correlation between the words in the filename created by the participant, and the words in the file that filename represents.

	M	*SD*
Different	0.32	0.03
None	0.33	0.05
Same	0.34	0.03
Self	0.35	0.03

path.file.similarity: correlation between the words in the full path in the hierarchy (excluding the filename) one would follow to reach a particular file, and the words in the file

	M	*SD*
Different	0.10	0.04
None	0.11	0.03
Same	0.10	0.04
Self	0.10	0.04

user.label.agreement: a *comparison measure*. This measure represents the percent of time two users chose the same words to represent the same file. For each file in the two hierarchies being compared, I counted the number of words that were the same in the full path to the file in both hierarchies, and divided that by the total number of distinct words from both hierarchies. I used the words from the full path, not just the file label, thinking that the full path is important when someone is navigating a hierarchy. I also did stemming and stop word removal; looked at the words individually, order isn't important.

G.3 Semantic Measures

dist.matrix.corr: a *comparison measure*. This measure represents the correlation between two distance matrices, each matrix representing one file-by-file hierarchy. The cells in the distance matrix contain the number of steps necessary to navigate from one file to the other in the hierarchy. This measure captures similarities between two participants' groupings of the files used in the experiment, without providing a qualitative assessment of the "rules" by which files seem to "go together" in a given hierarchy.

Appendix H

Regression Control Variables and Predictors

H.1 Controls and Predictors

The dependent variable in the model is $\mathrm{total.clicks}$, the total number of clicks (consisting of all folder open, folder close, and file view events) to locate the target file. What follows is a complete list of all the regressors included in the models presented in Chapters 4 and 5

Control variables

- $\mathrm{shortest.path}$: For each search task, the depth in the hierarchy of the target file, i.e., the absolute minimum number of clicks to find the target
- $\mathrm{consumer.id}$: Because each person experienced all types of search tasks, the model includes a fixed effects control for individual differences
- $\mathrm{hierarchy.id}$: The ID number of the hierarchy being searched; this is a fixed-effects control for differences between hierarchies
- $\mathrm{target.id}$: The search target file for a particular task. Some files may naturally be easier to find than other files, so this is also a fixed-effects control designed to capture that variability

Experimental variables

- $\mathrm{imagined.audience}$: The *Imagined Audience* for whom the hierarchy was created
- $\mathrm{PA.Same}$: Are the *Producer* and *Imagined Audience* from the same community? Yes or No

- AC.Same: Are the *Imagined Audience* and *Consumer* from the same community? Yes or No
- imagined.audience * AC.Same: Two-way interaction
- PA.Same * AC.Same: Two-way interaction

Vocabulary measures

- user.label.agreement: Comparing the hierarchy created by the consumer and the tree being searched, the agreement between the labels for all the files
- label.file.similarity: Vector correlation between the path (without the filename) and the target file in the PRODUCER's hierarchy, i.e., the hierarchy being searched
- path.file.similarity: Vector correlation between the filename and the target file in the PRODUCER's hierarchy, i.e., the hierarchy being searched

Topology measure

- average.path.length: The average number of steps from any file in a hierarchy to any other file, used as an indication of the complexity of the hierarchy; for example, a hierarchy with files grouped into only two folders at the same level has a lower average.path.length than a hierarchy with 4 or 5 levels and fewer files per folder

Semantic measure

- dist.matrix.corr: The correlation between the distance matrices for the hierarchy created by the consumer, and the hierarchy being searched in a given search task. a distance matrix for a given hierarchy consists of a file x file matrix with 33^2 cells; each cell contains the number of steps in the hierarchy to navigate from one file to the other file. This representation encodes information about how many files are close together in both hierarchies being compared.

H.2 Models

Model One: Hypothesis Testing

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{imagined.audience},\\ &\text{PA.Same}, \text{AC.Same}, \text{imagined.audience} * \text{AC.Same},\\ &\text{PA.Same} * \text{AC.same}, \text{consumer.id})\end{aligned}$$

Model Two: Hierarchy Measures Without PA.Same

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{dist.matrix.corr},\\ &\text{label.file.similarity}, \text{path.file.similarity}, \text{user.label.agreement},\\ &\text{PA.same}, \text{consumer.id})\end{aligned}$$

Model Three: Hierarchy Measures Without "Information Scent"

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{dist.matrix.cor},\\ &\text{user.label.agreement}, \text{PA.same}, \text{consumer.id},\end{aligned}$$

Model Four: Best Fit Hypothesis

$$\begin{aligned}\log(\text{total.clicks}) = f(&\text{shortest.path}, \text{average.path.length}, \text{label.file.similarity},\\ &\text{path.file.similarity}, \text{user.label.agreement},\\ &\text{imagined.audience} * \text{AC.Same}, \text{consumer.id})\end{aligned}$$

Model Five: Atheoretical Best Fit

$$\log(\text{total.clicks}) = f(\text{hierarchy.id}, \text{consumer.id}, \text{target.id})$$

Appendix I

Complete Regression Output

$$\log(\text{total.clicks}) = f(\text{shortest.path}, \text{average.path.length}, \text{imagined.audience}, \text{PA.Same}, \text{AC.Same}, \text{imagined.audience} * \text{AC.Same}, \text{PA.Same} * \text{AC.same}, \text{consumer.id})$$

Table I.1 Model (6.1) on page 85: Experiment IV's and Controls

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.7441691	0.2436197	3.0546	0.002253	*
shortest.path	0.1894845	0.0480749	3.9414	0.00008099	**
avg.path	0.2290423	0.0559147	4.0963	0.00004198	**
producer.audienceMSI	-0.0122437	0.1298375	-0.0943	0.924871	
producer.audienceself	0.1341051	0.1142022	1.1743	0.240284	
PA.sameY	-0.1833166	0.0889993	-2.0598	0.039422	
AC.sameY	-0.0619729	0.1501452	-0.4128	0.679787	
consumer.id19	0.5020840	0.2153444	2.3315	0.019725	
consumer.id21	0.0864202	0.1838994	0.4699	0.638404	
consumer.id25	0.2647506	0.2191524	1.2081	0.227022	
consumer.id30	-0.2806764	0.1841299	-1.5243	0.127424	
consumer.id31	0.0309832	0.2702935	0.1146	0.908740	
consumer.id40	0.3556323	0.2323630	1.5305	0.125892	
consumer.id41	0.3872864	0.2709153	1.4295	0.152847	
consumer.id44	-0.1603675	0.1937301	-0.8278	0.407790	
consumer.id46	-0.5751703	0.2188490	-2.6282	0.008585	*
consumer.id47	0.0269166	0.2002753	0.1344	0.893088	
consumer.id48	-0.1607851	0.1837478	-0.8750	0.381557	
consumer.id52	0.3836857	0.2629791	1.4590	0.144566	
consumer.id53	-0.0051575	0.1892786	-0.0272	0.978262	
consumer.id55	0.2030988	0.2404145	0.8448	0.398230	

Table I.1 Model (6.1) on page 85: Experiment IV's and Controls

	Estimate	Std. Error	z value	Pr(>|z|)	
consumer.id57	-0.6029202	0.1907502	-3.1608	0.001573	*
consumer.id59	0.3608897	0.2166039	1.6661	0.095688	.
consumer.id61	0.3553172	0.2398030	1.4817	0.138419	
consumer.id65	-0.0529149	0.2219205	-0.2384	0.811539	
consumer.id67	0.3349510	0.2080753	1.6098	0.107451	
consumer.id68	-0.0207004	0.1771417	-0.1169	0.906973	
consumer.id73	-0.3353683	0.2335533	-1.4359	0.151020	
consumer.id78	-0.3373867	0.2000720	-1.6863	0.091733	.
consumer.id80	0.2007520	0.2747253	0.7307	0.464940	
consumer.id81	0.0372311	0.2459226	0.1514	0.879665	
consumer.id82	0.3169686	0.3245734	0.9766	0.328782	
consumer.id84	-0.3011036	0.2241225	-1.3435	0.179117	
consumer.id90	0.1804483	0.2547787	0.7083	0.478787	
consumer.id92	-0.2762785	0.2100786	-1.3151	0.188470	
consumer.id93	-0.4518723	0.2045906	-2.2087	0.027198	
consumer.id94	-0.4656629	0.1861186	-2.5020	0.012350	
consumer.id96	0.3570702	0.2842677	1.2561	0.209078	
consumer.id102	0.0401744	0.1807045	0.2223	0.824064	
consumer.id106	0.3342013	0.2775291	1.2042	0.228511	
consumer.id107	-0.0596391	0.2349561	-0.2538	0.799626	
consumer.id108	0.1604993	0.2200114	0.7295	0.465693	
consumer.id109	0.0533584	0.2446139	0.2181	0.827325	
consumer.id110	0.2738068	0.2792542	0.9805	0.326843	
consumer.id111	0.2160385	0.2494311	0.8661	0.386422	
consumer.id117	0.3225589	0.2313734	1.3941	0.163286	
consumer.id120	-0.0358192	0.2614629	-0.1370	0.891034	
consumer.id121	0.2126729	0.2086220	1.0194	0.308005	
consumer.id125	0.0483658	0.2775854	0.1742	0.861679	
consumer.id131	0.3323884	0.2503243	1.3278	0.184234	
consumer.id133	0.5677289	0.2615740	2.1704	0.029974	
consumer.id135	-0.0103551	0.2582737	-0.0401	0.968018	
consumer.id140	0.2658493	0.2525689	1.0526	0.292533	
consumer.id143	-0.2876789	0.2413735	-1.1918	0.233323	
producer.audienceMSI:AC.sameY	0.0956274	0.2256147	0.4239	0.671673	
producer.audienceself:AC.sameY	-0.1670525	0.1913545	-0.8730	0.382663	
PA.sameY:AC.sameY	0.0370203	0.1254607	0.2951	0.767937	

* $p < .05$; ** $p < .01$; *** $p < .001$

$$\log(\text{total.clicks}) = f(\text{shortest.path}, \text{average.path.length}, \text{dist.matrix.cor}, \text{label.file.similarity}, \text{path.file.similarity}, \text{user.label.agreement}, \text{consumer.id})$$

Table I.2 Model (6.2) on page 85: Hierarchy Measures Only

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.6912429	0.2296676	3.0098	0.0026146	**
shortest.path	0.1695873	0.0476363	3.5600	0.0003708	***
avg.path	0.2787285	0.0551657	5.0526	0.0000004359	***
dist.cor	-0.5564219	0.2509873	-2.2169	0.0266277	*
filename.cor	-0.2402898	0.1281368	-1.8753	0.0607571	.
path.only.cor	-0.6111524	0.1849353	-3.3047	0.0009508	***
mean.label.iua	0.7487410	0.4430454	1.6900	0.0910304	.
consumer.id19	0.4004838	0.2061411	1.9428	0.0520445	.
consumer.id21	0.0355472	0.1754315	0.2026	0.8394266	
consumer.id25	0.2903963	0.2146428	1.3529	0.1760785	
consumer.id30	-0.3012769	0.1751749	-1.7199	0.0854572	.
consumer.id31	-0.0495067	0.2571468	-0.1925	0.8473325	
consumer.id40	0.3438940	0.2139312	1.6075	0.1079452	
consumer.id41	0.4452008	0.2993727	1.4871	0.1369852	
consumer.id44	-0.2108225	0.1873265	-1.1254	0.2604078	
consumer.id46	-0.6005450	0.2171106	-2.7661	0.0056735	**
consumer.id47	0.0361732	0.1965774	0.1840	0.8540016	
consumer.id48	-0.1707000	0.1802068	-0.9472	0.3435139	
consumer.id52	0.3507918	0.2339138	1.4997	0.1337017	
consumer.id53	-0.1002935	0.1938590	-0.5174	0.6049097	
consumer.id55	0.1404629	0.2211760	0.6351	0.5253809	
consumer.id57	-0.5994461	0.1835642	-3.2656	0.0010923	**
consumer.id59	0.3764574	0.2061053	1.8265	0.0677705	.
consumer.id61	0.3363090	0.2359417	1.4254	0.1540443	
consumer.id65	-0.1083205	0.2147738	-0.5043	0.6140175	
consumer.id67	0.3241429	0.2031731	1.5954	0.1106222	
consumer.id68	-0.0377812	0.1732215	-0.2181	0.8273439	
consumer.id73	-0.4382743	0.2269195	-1.9314	0.0534325	.
consumer.id78	-0.4081066	0.1804957	-2.2610	0.0237573	*
consumer.id80	0.0830927	0.2401133	0.3461	0.7293004	
consumer.id81	0.0087686	0.2273393	0.0386	0.9692328	
consumer.id82	0.3514661	0.2914145	1.2061	0.2277908	
consumer.id84	-0.3182209	0.2010824	-1.5825	0.1135263	
consumer.id90	0.1414628	0.2412260	0.5864	0.5575846	

Table I.2 Model (6.2) on page 85: Hierarchy Measures Only

	Estimate	Std. Error	z value	Pr(>|z|)	
consumer.id92	-0.3548204	0.1988475	-1.7844	0.0743612	.
consumer.id93	-0.5257058	0.1879405	-2.7972	0.0051549	**
consumer.id94	-0.4690762	0.1674238	-2.8017	0.0050830	**
consumer.id96	0.2779303	0.2668505	1.0415	0.2976341	
consumer.id102	0.0511226	0.1828163	0.2796	0.7797543	
consumer.id106	0.2062147	0.2313594	0.8913	0.3727589	
consumer.id107	-0.1431811	0.2067577	-0.6925	0.4886190	
consumer.id108	0.0238142	0.2167623	0.1099	0.9125177	
consumer.id109	-0.0108800	0.2251737	-0.0483	0.9614627	
consumer.id110	0.1758143	0.2620759	0.6709	0.5023144	
consumer.id111	0.1152417	0.2211005	0.5212	0.6022145	
consumer.id117	0.2946803	0.2022470	1.4570	0.1451077	
consumer.id120	-0.1051477	0.2519496	-0.4173	0.6764326	
consumer.id121	0.2199725	0.2134292	1.0307	0.3027013	
consumer.id125	-0.0062593	0.2674306	-0.0234	0.9813271	
consumer.id131	0.3499166	0.2334175	1.4991	0.1338473	
consumer.id133	0.5365396	0.2519125	2.1299	0.0331827	*
consumer.id135	-0.0986710	0.2272194	-0.4343	0.6641038	
consumer.id140	0.2393710	0.2386668	1.0030	0.3158846	
consumer.id143	-0.3672213	0.2396656	-1.5322	0.1254673	

* $p < .05$; ** $p < .01$; *** $p < .001$

$$\log(\text{total.clicks}) = f(\text{shortest.path}, \text{average.path.length}, \text{dist.matrix.cor},$$
$$\text{user.label.agreement}, \text{PA.same}, \text{consumer.id},$$

Table I.3 Model (6.3) on page 85: IV's and Hierarchy Measures w/o "Information Scent"

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.7647053	0.2465437	3.1017	0.0019241	**
shortest.path	0.1877800	0.0485105	3.8709	0.0001084	***
avg.path	0.2456176	0.0569253	4.3147	0.00001598	***
dist.cor	-0.5533331	0.2597068	-2.1306	0.0331216	*
mean.label.iua	0.6506483	0.4510232	1.4426	0.1491318	
PA.sameY	-0.1422912	0.0549423	-2.5898	0.0096023	**
consumer.id19	0.3945948	0.2202674	1.7914	0.0732235	.
consumer.id21	0.0385699	0.1860472	0.2073	0.8357657	
consumer.id25	0.2212758	0.2185230	1.0126	0.3112526	
consumer.id30	-0.3131029	0.1831794	-1.7093	0.0874011	.
consumer.id31	-0.0550567	0.2621501	-0.2100	0.8336522	
consumer.id40	0.3450077	0.2303745	1.4976	0.1342385	
consumer.id41	0.3627555	0.2772351	1.3085	0.1907119	
consumer.id44	-0.2064789	0.2009062	-1.0277	0.3040732	
consumer.id46	-0.6103557	0.2249055	-2.7138	0.0066510	**
consumer.id47	-0.0063003	0.2050451	-0.0307	0.9754877	
consumer.id48	-0.1741664	0.1852970	-0.9399	0.3472530	
consumer.id52	0.3428291	0.2495635	1.3737	0.1695303	
consumer.id53	-0.1133889	0.2009079	-0.5644	0.5724937	
consumer.id55	0.1654324	0.2382626	0.6943	0.4874765	
consumer.id57	-0.6125128	0.1963576	-3.1194	0.0018124	**
consumer.id59	0.3370066	0.2142512	1.5730	0.1157302	
consumer.id61	0.3205404	0.2402920	1.3340	0.1822164	
consumer.id65	-0.1043515	0.2270127	-0.4597	0.6457513	
consumer.id67	0.2989727	0.2132866	1.4017	0.1609924	
consumer.id68	-0.0533377	0.1844708	-0.2891	0.7724749	
consumer.id73	-0.4251747	0.2267879	-1.8748	0.0608247	.
consumer.id78	-0.4213120	0.1864096	-2.2601	0.0238125	*
consumer.id80	0.1037824	0.2512090	0.4131	0.6795103	
consumer.id81	-0.0365178	0.2314678	-0.1578	0.8746410	
consumer.id82	0.3211995	0.3042603	1.0557	0.2911175	
consumer.id84	-0.3540346	0.2058084	-1.7202	0.0853935	.
consumer.id90	0.1256452	0.2507399	0.5011	0.6163023 v	
consumer.id92	-0.3840003	0.2058395	-1.8655	0.0621067	.
consumer.id93	-0.5486589	0.1950139	-2.8134	0.0049015	**

Table I.3 IV's and Hierarchy Measures w/o "Information Scent"

	Estimate	Std. Error	z value	Pr(>|z|)	
consumer.id94	-0.5314952	0.1758804	-3.0219	0.0025118	**
consumer.id96	0.2622942	0.2674902	0.9806	0.3268024	
consumer.id102	0.0217346	0.1826904	0.1190	0.9052995	
consumer.id106	0.2288715	0.2436390	0.9394	0.3475318	
consumer.id107	-0.1398140	0.2218161	-0.6303	0.5284885	
consumer.id108	0.0095844	0.2250488	0.0426	0.9660300	
consumer.id109	-0.0140469	0.2457814	-0.0572	0.9544241	
consumer.id110	0.1644358	0.2699727	0.6091	0.5424695	
consumer.id111	0.1329331	0.2319352	0.5731	0.5665448	
consumer.id117	0.2703197	0.2121697	1.2741	0.2026376	
consumer.id120	-0.0992268	0.2674888	-0.3710	0.7106697	
consumer.id121	0.1719974	0.2116412	0.8127	0.4163993	
consumer.id125	-0.0047080	0.2665058	-0.0177	0.9859055	
consumer.id131	0.3143630	0.2404208	1.3076	0.1910250	
consumer.id133	0.5353524	0.2623108	2.0409	0.0412599	*
consumer.id135	-0.1285736	0.2374786	-0.5414	0.5882242	
consumer.id140	0.2195805	0.2438722	0.9004	0.3679118	
consumer.id143	-0.3659457	0.2413399	-1.5163	0.1294414	

* $p < .05$; ** $p < .01$; *** $p < .001$

$$\log(\text{total.clicks}) = f(\text{shortest.path}, \text{average.path.length}, \text{label.file.similarity}, \text{path.file.similarity}, \text{user.label.agreement}, \text{imagined.audience} * \text{AC.Same}, \text{consumer.id}) \quad \text{(I.1)}$$

Table I.4 Model (6.4) on page 85: All Variables

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.7474590	0.2382922	3.1367	0.0017084	**
shortest.path	0.1862208	0.0476048	3.9118	0.00009161	***
avg.path	0.2386771	0.0551987	4.3240	0.00001532	***
filename.cor	-0.3019658	0.1346318	-2.2429	0.0249032	*
path.only.cor	-0.6769937	0.1876218	-3.6083	0.0003082	***
mean.label.iua	0.5940704	0.4451190	1.3346	0.1819966	
producer.audienceMSI	-0.0534143	0.1291211	-0.4137	0.6791117	
producer.audienceself	-0.0422671	0.1070107	-0.3950	0.6928580	
AC.sameY	-0.0559168	0.1314914	-0.4253	0.6706540	
consumer.id19	0.4917439	0.2022010	2.4320	0.0150175	*
consumer.id21	0.0686250	0.1729160	0.3969	0.6914642	
consumer.id25	0.3357997	0.2166323	1.5501	0.1211197	
consumer.id30	-0.2693453	0.1714139	-1.5713	0.1161095	
consumer.id31	0.0231330	0.2579102	0.0897	0.9285304	
consumer.id40	0.3521301	0.2162915	1.6280	0.1035175	
consumer.id41	0.4821770	0.2955148	1.6317	0.1027531	
consumer.id44	-0.1453100	0.1847588	-0.7865	0.4315836	
consumer.id46	-0.5413386	0.2189024	-2.4730	0.0133996	*
consumer.id47	0.0701330	0.1944259	0.3607	0.7183101	
consumer.id48	-0.1627799	0.1775519	-0.9168	0.3592464	
consumer.id52	0.4040568	0.2515878	1.6060	0.1082680	
consumer.id53	0.0203151	0.1824662	0.1113	0.9113496	
consumer.id55	0.1822991	0.2271915	0.8024	0.4223199	
consumer.id57	-0.5739271	0.1798900	-3.1904	0.0014206	**
consumer.id59	0.3990797	0.2041282	1.9550	0.0505778	.
consumer.id61	0.3590605	0.2308128	1.5556	0.1197949	
consumer.id65	-0.0654267	0.2109811	-0.3101	0.7564796	
consumer.id67	0.3657479	0.1980083	1.8471	0.0647276	.
consumer.id68	0.0083153	0.1663500	0.0500	0.9601328	
consumer.id73	-0.3580941	0.2317418	-1.5452	0.1222909	
consumer.id78	-0.3092894	0.1935446	-1.5980	0.1100370	
consumer.id80	0.1764848	0.2709961	0.6512	0.5148885	
consumer.id81	0.0913866	0.2421728	0.3774	0.7059053	
consumer.id82	0.3879423	0.3185602	1.2178	0.2233003	

Table I.4 Model (6.4) on page 85: All Variables

	Estimate	Std. Error	z value	Pr(>|z|)	
consumer.id84	-0.2501731	0.2206831	-1.1336	0.2569495	
consumer.id90	0.1990229	0.2404170	0.8278	0.4077703	
consumer.id92	-0.2563033	0.2017640	-1.2703	0.2039735	
consumer.id93	-0.4165581	0.1962713	-2.1224	0.0338077	*
consumer.id94	-0.3616302	0.1828654	-1.9776	0.0479765	*
consumer.id96	0.3829936	0.2795251	1.3702	0.1706375	
consumer.id102	0.0978662	0.1810633	0.5405	0.5888465	
consumer.id106	0.3273643	0.2751036	1.1900	0.2340592	
consumer.id107	-0.0524206	0.2251888	-0.2328	0.8159282	
consumer.id108	0.1996011	0.2140734	0.9324	0.3511321	
consumer.id109	0.0785422	0.2237896	0.3510	0.7256151	
consumer.id110	0.2768716	0.2660891	1.0405	0.2980974	
consumer.id111	0.1985709	0.2414732	0.8223	0.4108884	
consumer.id117	0.3680476	0.2228322	1.6517	0.0985997	.
consumer.id120	-0.0366756	0.2415141	-0.1519	0.8792999	
consumer.id121	0.2530585	0.2069933	1.2225	0.2215018	
consumer.id125	0.0402907	0.2689943	0.1498	0.8809360	
consumer.id131	0.3824036	0.2464900	1.5514	0.1208069	
consumer.id133	0.5687888	0.2559873	2.2219	0.0262872	*
consumer.id135	0.0201720	0.2416399	0.0835	0.9334703	
consumer.id140	0.3198524	0.2523281	1.2676	0.2049390	
consumer.id143	-0.2870761	0.2354247	-1.2194	0.2226938	
producer.audienceMSI:AC.sameY	0.1077338	0.2229753	0.4832	0.6289786	
producer.audienceself:AC.sameY	-0.1431665	0.1819455	-0.7869	0.4313609	

* $p < .05$; ** $p < .01$; *** $p < .001$

$$\log(\text{total.clicks}) = f(\text{hierarchy.id}, \text{consumer.id}, \text{target.id}) \tag{I.2}$$

Table I.5 Model (6.5) on page 86: Atheoretical Best Fit

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.212325	0.212786	10.397	< 0.000000	***
producer.id19	-0.142212	0.291240	-0.488	0.625337	
producer.id21	-0.146964	0.253698	-0.579	0.562397	
producer.id22	0.199821	0.227166	0.880	0.379062	
producer.id25	0.103312	0.230563	0.448	0.654092	
producer.id29	0.009165	0.239036	0.038	0.969416	
producer.id30	0.153964	0.223073	0.690	0.490070	
producer.id31	-0.360262	0.219480	-1.641	0.100708	
producer.id32	-0.045933	0.241269	-0.190	0.849011	
producer.id35	-0.032421	0.258625	-0.125	0.900238	
producer.id38	-0.088696	0.241482	-0.367	0.713398	
producer.id40	-0.029788	0.235144	-0.127	0.899194	
producer.id44	0.281636	0.226125	1.245	0.212952	
producer.id45	-0.244139	0.231973	-1.052	0.292596	
producer.id47	-0.229204	0.245356	-0.934	0.350216	
producer.id48	0.204186	0.243977	0.837	0.402646	
producer.id52	-0.108933	0.238326	-0.457	0.647618	
producer.id53	0.750231	0.202339	3.708	0.000209	***
producer.id54	0.127117	0.212462	0.598	0.549636	
producer.id57	0.093741	0.207721	0.451	0.651786	
producer.id58	-0.498452	0.261809	-1.904	0.056926	.
producer.id59	-0.270011	0.262319	-1.029	0.303329	
producer.id62	-0.409556	0.228232	-1.794	0.072738	.
producer.id64	-0.403450	0.257062	-1.569	0.116539	
producer.id65	-0.045856	0.237276	-0.193	0.846753	
producer.id66	0.002895	0.226461	0.013	0.989801	
producer.id67	0.078104	0.259584	0.301	0.763504	
producer.id69	-0.247733	0.238485	-1.039	0.298909	
producer.id70	-0.201008	0.244846	-0.821	0.411670	
producer.id72	0.681820	0.212293	3.212	0.001320	**
producer.id73	-0.551909	0.250741	-2.201	0.027728	*
producer.id76	0.385048	0.254806	1.511	0.130752	
producer.id78	0.265478	0.222571	1.193	0.232955	
producer.id80	0.337173	0.244920	1.377	0.168616	
producer.id81	-0.844221	0.246894	-3.419	0.000628	***
producer.id82	0.506742	0.226219	2.240	0.025088	*

Table I.5 Model (6.5) on page 86: Atheoretical Best Fit

	Estimate	Std. Error	z value	Pr(>|z|)	
producer.id89	-0.190383	0.241110	-0.790	0.429755	
producer.id96	0.160782	0.238991	0.673	0.501104	
producer.id98	-0.118814	0.258168	-0.460	0.645360	
producer.id99	-0.495568	0.242012	-2.048	0.040589	*
producer.id101	0.594765	0.203275	2.926	0.003434	**
producer.id102	-0.786331	0.243720	-3.226	0.001254	**
producer.id106	-0.105317	0.264584	-0.398	0.690595	
producer.id107	-0.047065	0.253335	-0.186	0.852616	
producer.id108	-1.026757	0.255059	-4.026	0.0000568	***
producer.id109	0.079002	0.246052	0.321	0.748151	
producer.id110	-0.151166	0.242847	-0.622	0.533631	
producer.id111	-0.858090	0.250268	-3.429	0.000607	***
producer.id112	-0.977219	0.238796	-4.092	0.0000427	***
producer.id114	-0.870896	0.253986	-3.429	0.000606	***
producer.id117	-0.029610	0.246780	-0.120	0.904494	
producer.id118	-0.794110	0.314461	-2.525	0.011560	*
producer.id121	-0.677352	0.250103	-2.708	0.006763	**
producer.id123	0.264724	0.233492	1.134	0.256895	
producer.id125	-0.430297	0.256754	-1.676	0.093756	.
producer.id126	0.140674	0.231975	0.606	0.544236	
producer.id129	-0.210313	0.258407	-0.814	0.415712	
producer.id130	-0.504974	0.283969	-1.778	0.075359	.
producer.id131	-0.048338	0.235790	-0.205	0.837570	
producer.id133	-0.297056	0.273836	-1.085	0.278013	
producer.id135	-0.127509	0.233219	-0.547	0.584561	
producer.id137	-0.073054	0.224442	-0.325	0.744810	
producer.id138	-0.039848	0.249355	-0.160	0.873035	
producer.id140	-0.245849	0.234226	-1.050	0.293892	
consumer.id19	0.457924	0.184486	2.482	0.013059	*
consumer.id21	-0.024724	0.188942	-0.131	0.895889	
consumer.id25	0.160384	0.185979	0.862	0.388480	
consumer.id30	-0.374044	0.197476	-1.894	0.058209	.
consumer.id31	-0.113771	0.191007	-0.596	0.551420	
consumer.id40	0.188853	0.185359	1.019	0.308275	
consumer.id41	0.274990	0.188360	1.460	0.144313	
consumer.id44	-0.284772	0.191424	-1.488	0.136844	
consumer.id46	-0.615477	0.200499	-3.070	0.002143	**
consumer.id47	0.010177	0.195661	0.052	0.958519	
consumer.id48	-0.220917	0.191362	-1.154	0.248317	
consumer.id52	0.248412	0.190056	1.307	0.191197	
consumer.id53	-0.039801	0.191625	-0.208	0.835459	

Table I.5 Model (6.5) on page 86: Atheoretical Best Fit

	Estimate	Std. Error	z value	Pr(>|z|)	
consumer.id55	0.079462	0.186534	0.426	0.670115	
consumer.id57	-0.739681	0.203325	-3.638	0.000275	***
consumer.id59	0.211488	0.185287	1.141	0.253700	
consumer.id61	0.177727	0.186803	0.951	0.341394	
consumer.id65	-0.117596	0.194533	-0.605	0.545509	
consumer.id67	0.187246	0.192007	0.975	0.329457	
consumer.id68	-0.064318	0.190088	-0.338	0.735092	
consumer.id73	-0.572576	0.198567	-2.884	0.003932	**
consumer.id78	-0.458239	0.195281	-2.347	0.018948	*
consumer.id80	0.016513	0.186799	0.088	0.929560	
consumer.id81	-0.088352	0.193822	-0.456	0.648505	
consumer.id82	0.206742	0.191148	1.082	0.279439	
consumer.id84	-0.404251	0.194554	-2.078	0.037725	*
consumer.id90	-0.052532	0.187515	-0.280	0.779366	
consumer.id92	-0.286286	0.193871	-1.477	0.139760	
consumer.id93	-0.684487	0.200695	-3.411	0.000648	***
consumer.id94	-0.531696	0.196975	-2.699	0.006948	**
consumer.id96	0.075743	0.187690	0.404	0.686540	
consumer.id102	-0.119004	0.189328	-0.629	0.529636	
consumer.id106	0.234380	0.190359	1.231	0.218230	
consumer.id107	-0.259492	0.191175	-1.357	0.174670	
consumer.id108	0.041114	0.189362	0.217	0.828116	
consumer.id109	-0.182274	0.191297	-0.953	0.340676	
consumer.id110	0.017316	0.190621	0.091	0.927621	
consumer.id111	0.044754	0.187295	0.239	0.811143	
consumer.id117	0.180018	0.186121	0.967	0.333437	
consumer.id120	-0.227881	0.191039	-1.193	0.232929	
consumer.id121	0.202846	0.193385	1.049	0.294214	
consumer.id125	-0.133265	0.192227	-0.693	0.488143	
consumer.id131	0.222401	0.193218	1.151	0.249718	
consumer.id133	0.403656	0.182509	2.212	0.026987	*
consumer.id135	-0.236232	0.194682	-1.213	0.224967	
consumer.id140	0.089090	0.185612	0.480	0.631244	
consumer.id143	-0.394919	0.194647	-2.029	0.042468	*
target.doc.id150	-0.219080	0.132641	-1.652	0.098601	.
target.doc.id151	0.349922	0.134631	2.599	0.009346	**
target.doc.id152	-0.010287	0.134381	-0.077	0.938979	
target.doc.id153	-0.111951	0.137295	-0.815	0.414843	
target.doc.id154	0.274250	0.139985	1.959	0.050096	.
target.doc.id155	-0.210380	0.136159	-1.545	0.122322	
target.doc.id156	-0.293853	0.139524	-2.106	0.035195	*

Table I.5 Model (6.5) on page 86: Atheoretical Best Fit

	Estimate	Std. Error	z value	Pr(>|z|)	
target.doc.id157	-0.493595	0.142053	-3.475	0.000511	***
target.doc.id158	0.065842	0.133393	0.494	0.621594	
target.doc.id159	-0.060590	0.132281	-0.458	0.646923	
target.doc.id163	0.058787	0.147988	0.397	0.691189	
target.doc.id165	-0.002619	0.143986	-0.018	0.985485	
target.doc.id167	-0.237375	0.135728	-1.749	0.080307	.
target.doc.id168	0.310152	0.140303	2.211	0.027065	*
target.doc.id169	0.296086	0.139299	2.126	0.033541	*
target.doc.id170	-0.361165	0.147876	-2.442	0.014592	*
target.doc.id171	0.343383	0.137096	2.505	0.012256	*
target.doc.id172	0.241057	0.141679	1.701	0.088862	.

* $p < .05$; ** $p < .01$; *** $p < .001$

Appendix J

R Code for Model Comparison and Goodness of Fit

This appendix contains R code written for the model selection and goodness of fit analysis. The formulas are based on the following references, cited in the comments where appropriate: Anderson (2008); Burnham and Anderson (2004); Fridstrøm, Ifver, Ingebrightsen, Kulmala, and Thomsen (1995); Menard (2002); Wagenmakers and Farrell (2004)

```
# model_selection function takes a list of negbin model objects, and outputs a table
    containing model comparison statistics

model_selection <- function(x,d) {
   table.df <- data.frame(row.names=1:length(x), model=NA, aic=NA, aicc=NA, qaicc=NA,
       delta.aic=NA, relative.lik.aic=NA, aic.weight=NA, aic.ratio=NA, bic=NA,
       delta.bic=NA, relative.lik.bic=NA, bic.weight=NA, bic.ratio=rep(NA, length(x)))

   for (i in 1:length(x)) {
      table.df$model[i] <- names(x)[i]
      K <- length(x[[i]]$coefficients)
      c.hat <- x[[i]]$deviance / x[[i]]$df.residual

      # AIC = -2(log-likelihood) + 2K (where K is the number of coefficients, including
          the intercept, plus 1 for the overdispersion estimate)
      table.df$aic[i] <- (logLik(x[[i]])[1] * -2) + (2 * K)

      # AICc -- an adjustment for small sample size
      table.df$aicc[i] <- table.df$aic[i] + ((2*K)*(K+1) / (length(d$total)-K-1))

      #QAICc -- an adjustment for overdispersion AND small sample size (this is the
          appropriate one for the finding experiment data)
      table.df$qaicc[i] <- (-1*(logLik(x[[i]])[1] * 2) / c.hat) + (2*K) + ((2*K)*(K+1) /
          (length(d$total)-K-1))

      # BIC -- an alternative to AIC
      table.df$bic[i] <- (logLik(x[[i]])[1] * -2) + length(x[[i]]$coefficients) *
          log(length(d$total))
   }

   for (i in 1:length(x)) {
      # delta.aic  -- the difference between the AICs of two different models
      table.df$delta.aic[i] <- table.df$qaicc[i] - min(table.df$qaicc)
      table.df$delta.bic[i] <- table.df$bic[i] - min(table.df$bic)

      # the relative likelihood of a model, given the data and the set of models under
          consideration: exp ( - 1/2 * delta-sub-i) -- "useful in making inferences
```

```
            concerning the relative strength of evidence for each of the models in the set"
            p74 (Burnham & Anderson, 2004). The odds for the ith model actually being the
            K-L best model... the Akaike weights are an effective way to scale and
            interpret the delta-sub-i values
         table.df$relative.lik.aic[i] <- exp(-1/2 * table.df$delta.aic[i])
         table.df$relative.lik.bic[i] <- exp(-1/2 * table.df$delta.bic[i])
      }

      for (i in 1:length(x)) {
         # Akaike weight or w-sub-i = exp (- 1/2 * delta.aic ) / sum of exp (- 1/2 *
            delta.aic) for all the models under contention -- "A given Akaike weight is
            considered as the weight of evidence in favor of model i being the actual K-L
            best model for the situation at hand *given* that one of the R models must be
            the K-L best model of that set of R models.
         # the AIC/BIC weight is interpreted as the probability that the model in question
            is the "best" model, given the options (Wagenmakers & Farrell, 2004)
         table.df$aic.weight[i] <- table.df$relative.lik.aic[i] /
            sum(table.df$relative.lik.aic, na.rm=T)
         table.df$bic.weight[i] <- table.df$relative.lik.bic[i] /
            sum(table.df$relative.lik.bic, na.rm=T)
      }

      # evidence ratios (p77, Burnham & Anderson, 2004) = weight of one model / weight of
         another model + weight of first model. in other words, the relative likelihood of
         one model vs another model, given the data. Evidence ratios can be expressed as
         the "normalized probability that one model is preferred over another" (Wagenmakers
         & Farrell, 2004)
      for (i in 1:length(x)) {
         a = max(table.df$aic.weight)
         b = max(table.df$bic.weight)
         if (a != table.df$aic.weight[i]) { table.df$aic.ratio[i] <- table.df$aic.weight[i]
            / (table.df$aic.weight[i] + a) }
         if (b != table.df$bic.weight[i]) { table.df$bic.ratio[i] <- table.df$bic.weight[i]
            / (table.df$bic.weight[i] + b) }
      }

      table.df$relative.lik.aic <- format.pval(table.df$relative.lik.aic)
      table.df$aic.weight <- format.pval(table.df$aic.weight)
      table.df$aic.ratio <- format.pval(table.df$aic.ratio)

      table.df$relative.lik.bic <- format.pval(table.df$relative.lik.bic)
      table.df$bic.weight <- format.pval(table.df$bic.weight)
      table.df$bic.ratio <- format.pval(table.df$bic.ratio)

      colnames(table.df) <- c("model", "AIC", "AICc", "QAICc", "delta.QAICc",
          "relative.lik.QAICc", "QAICc.weight", "QAICc.ratio", "BIC", "delta.BIC",
          "relativelik.BIC", "BIC.weight", "BIC.ratio")

      return(table.df)
   }

   # model_fit function calculates a variety of goodness of fit measures

   model_fit <- function(m,d) {
      # test the model against the saturated model to get dev.p (the residual deviance is
         the one i want, or model$deviance). if significant, reject the null hypothesis
         that the full model is no different from the saturated model. want hight p-values
         (no significance)
      dev.p <- pchisq(m$deviance, m$df.residual, lower.tail=F)

      # Regress the dependent var with no IVs; just the intercept
      simple_fit <- glm(total ~ 1, data=d, family=negative.binomial(theta=m$theta))

      # Gm = difference between deviance of simple model and deviance of the model in
         question. if Gm is significant, reject the nulll hypothesis that the full model is
         no different from the simple model. want LOW p-values, this shows the model with
         predictors is better than the model without predictors
```

```
    D.null <- logLik(simple_fit)[1] * -2
    D.model <- logLik(m)[1] * -2
    Gm <- D.null - D.model # the model chi-square

    # R2 from Anderson, 2008, appendix A, p154
    R2A <- (1 - exp(-2/length(d$total) * (logLik(m)[1] - logLik(simple_fit)[1]))) / (1 -
        exp(2/length(d$total) * logLik(simple_fit)[1]))

    # Calculate the degrees of freedom for the GM goodness of fit test: this is the number
        of coefficients - 1
    Gm_df <- length(m$coefficients) - 1

    # Calculate the p-value for the Gm goodness of fit test
    Gm_p <- format.pval(pchisq(Gm, Gm_df, lower.tail=F))

    # R-squared, the coefficient of determination, calculated like it is for OLS (the
        correlation between the fitted and observed values)
    R_2 <- 1 - sum((d$total - m$fitted.values)^2) / sum((d$total - mean(d$total))^2)

    # calculate the R^2_L -- the percentage of the likelihood explained by the model
        (this is McFaddens pseudo R-square)
    # also called the Pseudo (McFadden) R-square = 1 - Loglik(model)/Loglik(null model),
        where the null model includes only the intercept
    # this ratio represents amount of improvement over the null model (essentially, a
        negative binomial distribution with an intercept); therefore, this isn't about
        absolute fit to the data (from
        http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm)
    R_2_L <- Gm / D.null

    # adjusted McFaddens pseudo R-square includes a penalty for including too many
        predictors (like adjusted R2 in OLS) -- also from
        http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm
    K <- length(m$coefficients) + 1
    adjusted.R_2_L <- 1 - (((m$twologlik / 2) - K) / (logLik(simple_fit)[1]))

    # P2 represents the most variance that can be explained in a perfectly specified and
        estimated Poisson model. the scaled R2p is the proportion of potentially
        explainable systematic variation that can be explained given the predictors
        (Fridstrm et. al, 1995)
    P2 <- 1 - sum(m$fitted.values) / sum((d$total - mean(d$total))^2)
    R2_p <- R_2 / P2

    return(list(Dm=m$deviance, Dmp=dev.p, Gm=Gm, Gmp=Gm_p, R2A=R2A, R2=R_2, R2L=R_2_L,
        adj.R2L=adjusted.R_2_L, R2P=R2_p))
}
```

References

Ackerman, M. S. (2000). The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Human-Computer Interaction*, *15*(2), 181 - 203.

Agresti, A. (2007). *An introduction to categorical data analysis* (Second ed.). Wiley.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*, 3-14.

Aloise, G., Castellano, M., Cogliani, L., Gill, J., Mak, M., Noel, J., et al. (2009, March 2). *NNSA and DOD need to more effectively manage the stockpile life extension program.* Retrieved August 10, 2009, from http://www.gao.gov/products/GAO-09-385

Anderson, D. R. (2008). *Model based inference in the life sciences*. New York, NY: Springer Science+Business Media, LLC.

Anthias, T., & Sankar, K. (2008). The network's new role. *ACM Queue*, *4*(4), 38-46.

Bae, S., Marshall, C. C., Meintanis, K., Zacchi, A., Hsieh, H., Moore, J. M., et al. (2006). Patterns of reading and organizing information in document triage. In *Proceedings of the ASIS&T 2006 Annual Meeting.*

Barr, J., & Cabrera, L. F. (2008). Ai gets a brain. *ACM Queue*, *4*(4), 24-29.

Barreau, D. (1995). Context as a factor in personal information management systems. *Journal of the American Society for Information Science*, *46*(5), 327-339.

Basili, V. R., & Zelkowitz, M. V. (2008). Empirical studies to build a science of computer science. *Communications of the ACM*, *50*(11), 33-37.

Bates, M. J. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, *49*(13), 1185-1205.

Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., & Grinter, R. (2005). Quality vs. quantity: Email-centric task-management and its relationship with overload. *Human-Computer Interaction*, *20*(1-2), 89-138.

Bergman, O., Beyth-Marom, R., Nachimas, R., Gradovitch, N., & Whittaker, S. (2008). Improved search engines and navigation preference in personal information management.

ACM Transactions on Information Systems, *26*(4), Article 20.

Berlin, L. M., Jeffries, R., O'Day, V. L., Paepcke, A., & Wharton, C. (1993). Where did you put it? Issues in the design and use of a group memory. In *CHI '93: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 23-30). Amsterdam, The Netherlands: ACM Press.

Blair, D. C., & Kimbrough, S. O. (2002). Exemplary documents: A foundation for information retrieval design. *Information Processing and Management*, *38*, 363-379.

Boardman, R., & Sasse, M. A. (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* (p. 583-590). Vienna, Austria.

Boh, W. F. (2007). Mechanisms for sharing knowledge in project-based organizations. *Information and Organization*, *17*, 27-58.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., et al. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, *24*(3), 127-135.

Boyd, A. (2005). *Organizations shift focus to information management: The role of documents in highly effective business processes.* Retrieved August 8, 2009, from http://www.edsf.org/includes/downloads/organization_shift_focus.pdf

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*(6), 1482-1493.

Bruce, H., Jones, W., & Dumais, S. (2004). Information behaviour that keeps found things found. *Information Research*, *10*(1).

Bullen, A. (2008). The 'long tale': using web 2.0 concepts to enhance digital collections. *Computers in Libraries*, *31*(5).

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research*, *33*, 261-304.

Byers, A. L., Allore, H., Gill, T. M., & Peduzzi, P. N. (2003). Application of negative binomial modeling for discrete outcomes: A case study in aging research. *Journal of Clinical Epidemiology*, *56*(6), 559 - 564.

Caldas, A., Schroeder, R., Mesch, G. S., & Dutton, W. H. (2008). Patterns of information search and access on the world wide web: Democratizing expertise or creating new hierarchies? *Journal of Computer-Mediated Communication*, *13*(4), 769-793.

Callow, M., Beardow, P., & Brittain, D. (2008). Big games, small screens. *ACM Queue*, *5*(7), 40-50.

Carmel, E., & Abbott, P. (2008). Why 'nearshore' means that distance matters. *Communications of the ACM*, *50*(10), 40-46.

Carroll, J. M. (1980). The role of context in creating names. *Discourse Processes*, *3*(1), 1-24.

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

Chalmers, M. (2003). Informatics, architecture and language. In K. Hook, D. Benyon, & A. J. Munro (Eds.), *Designing information spaces: The social navigation approach* (p. 315-342). London: Springer.

Chang, M. K., & Law, S. P. M. (2008). Factor structure for Young's Internet Addiction Test: A confirmatory study. *Computers in Human Behavior*, *24*(6), 2597-2619.

Chang, S.-J., & Rice, R. E. (1993). Browsing: A multidimensional framework. *Annual Review of Information Science and Technology*, *28*, 231-271.

Chantraine, Y., & Hupet, M. (1994). Efficiency of the addressee's contribution to the establishment of references: Comparing monologues with dialogues. *Cahier de psychologie cognitive (Current Psychology of Cognition)*, *13*(6), 777-796.

Cheshire, C., & Antin, J. (2008). The social psychological effects of feedback on the production of internet information pools. *Journal of Computer-Mediated Communication*, *13*(3), 705-727.

Chiou, J.-S., & Lee, J. (2008). What do they say about "Friends"? A cross-cultural study on internet discussion forum. *Computers in Human Behavior*, *24*(3), 1179-1195.

Civan, A., Jones, W., Klasnja, P., & Bruce, H. (2008). Better to organize personal information by folders or by tags? The devil is in the details. In *Proceedings of the ASIS&T 2008 Annual Meeting.*

Clark, H. H. (1996). Common ground. In *Using language.* Cambridge: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (p. 127-149). Washington DC: American Psychological Association.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62 - 81.

Cowan, R., David, P. A., & Foray, D. (2000). The explicit economics of knowledge codification and tacitness. *Industrial and Corporate Change*, *9*(2), 211-253.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed-method approaches* (Second ed.). Thousand Oaks, CA: SAGE Publications.

Crosby, S., & Brown, D. (2008). The virtualization reality. *ACM Queue*, *4*(10), 34-41.

Cutrell, E., Robbins, D., Dumais, S., & Sarin, R. (2006). Fast, flexible filtering with phlat. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 261-270). New York, NY, USA: ACM.

Cymru, T. (2008). Cybercrime: an epidemic. *ACM Queue*, *4*(9), 24-35.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, *8*(1), 19-30.

Dubberly, H. (2008). Design in the age of biology: shifting from a mechanical-object ethos to an organic-systems ethos. *interactions*, *15*(5), 35-41.

Elerath, J. (2008). Hard disk drives: the good, the bad, and the ugly. *ACM Queue*, *5*(6), 28-37.

Elsweiler, D., Ruthven, I., & Jones, C. (2007). Towards memory supporting personal information management tools. *Journal of the American Society for Information Science and Technology*, *58*(7), 924 - 946.

Erickson, T., & Kellogg, W. (2000). Social translucence: an approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction*, *7*(1), 59-83.

Feldman, S., Duhl, J., Marobella, J. R., & Crawford, A. (2005). *The hidden costs of information work.* Retrieved August 8, 2009, from http://factiva.com/collateral/files/whitepaper_IDC_hiddencosts_0405.pdf

Franzke, M. (1995). Turning research into practice: characteristics of display-based interaction. In *CHI '95: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 421-428). New York, NY, USA: ACM.

Fridstrøm, L., Ifver, J., Ingebrightsen, S., Kulmala, R., & Thomsen, L. K. (1995). Measuring the contribution of randomness, exposure, weather and daylight to the variation in road accident counts. *Accident Analysis and Prevention*, *27*(1), 1-20.

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1983). Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal*, *62*(6), 1753-1806.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Commun. ACM*, *30*(11), 964-971.

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, *25*(3), 203-219.

Garfinkel, S. L. (2008). Document & media exploitation. *ACM Queue*, *5*(7), 22-30.

Garrett, R. K., & Danziger, J. N. (2008). Disaffection or expected outcomes: Understanding personal internet use during work. *Journal of Computer-Mediated Communication*,

13(4), 937-958.

Geer, D. E. (2008). Playing for keeps. *ACM Queue*, *4*(9), 42-48.

Gelman, S., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183–209.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2006). The impact of delayed visual feedback on collaborative performance. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 1303-1312). New York, NY, USA: ACM.

Ghidey, W., Lesaffre, E., & Verbeke, G. (2008, June 18). A comparison of methods for estimating the random effects distribution of a linear mixed model. *Statistical Methods in Medical Research*, 1-26.

Goeke, R. J., & Faley, R. H. (2008). Leveraging the flexibility of your data warehouse. *Communications of the ACM*, *50*(10), 107-111.

Gordon, M. D. (1997). It's 10 am, do you know where your documents are? The nature and scope of information retrieval problems in business. *Information Processing & Management*, *33*(1), 107-122.

Greenberg, J., Crystal, A., Robertson, W. D., & Leadem, E. (2003). Iterative design of metadata creation tools for resource authors. In *2003 dublin core conference: Supporting communities of discourse and practice—metadata research and applications.* Seattle, Washington.

Grudin, J. (2006, 4-7 January). Enterprise knowledge management and emerging technologies. In *HICSS '06.*

Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). McGraw-Hill/Irwin.

Henning, M. (2008). API design matters. *ACM Queue*, *5*(4), 24-36.

Hertzum, M. (1999). Six roles of documents in professionals' work. In *ECSCW '99: Proceedings of the Sixth European Conference on Computer-Supported Cooperative Work.* Netherlands: Springer.

Hertzum, M., & Pejtersen, A. M. (2000). The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing & Management*, *36*(1), 761-778.

Huwe, T. K. (2008). The surprising impact of digital repositories. *Computers in Libraries*, *42*(3).

Isaacs, E., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*(1), 26-37.

Jacob, E. K. (1995). Communication and category structure: the communicative process as a constraint on the semantic representation of information. In B. H. Kwasnik, P. Smith, R. Fidel, & C. Beghtol (Eds.), *Advances in classification research, vol. 4.* Medford,

NJ: Information Today.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2008). Social phishing. *Communications of the ACM*, *50*(10), 94-100.

Jian, G., & Jeffres, L. (2006). Understanding employees' willingness to contribute to shared electronic databases: A three dimensional framework. *Communication Research*, *33*(4), 242-261.

Jones, W., Phuwanartnurak, A. J., Gill, R., & Bruce, H. (2005). Don't take my folders away! Organizing personal information to get things done. In *CHI '05: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 1505-1508). New York, NY, USA: ACM Press.

Kellogg, W. A., & Breen, T. J. (1987). Evaluating user and system models: Applying scaling techniques to problems in human-computer interaction. In *CHI '87: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 303-308). New York, NY, USA: ACM.

Keysar, B., & Henly, A. (2002). Speakers' overestimation of their effectiveness. *Psychological Science*, *13*(3), 207-212.

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: representations of others' knowledge in reference. *Social Cognition*, *9*(2-24).

Krauss, R. M., Vivekananthan, P., & Weinheimer, S. (1968). 'Inner speech' and 'External speech': Characteristics and communication effectiveness of socially and nonsocially encoded messages. *Journal of Personality and Social Psychology*, *9*(4), 295-300.

Kwasnik, B. (1989). How a personal document's intended use or purpose affects its classification in an office. In *Proceedings of the 12th annual international ACM SIGIR Conference on Research and Development in Information Retrieval.* Cambridge, Massachusetts, United States.

Lampe, C. A., Ellison, N., & Steinfield, C. (2007). A familiar face(book): profile elements as signals in an online social network. In *CHI '07: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 435-444). New York, NY, USA: ACM.

Lansdale, M. (1988). The psychology of personal information management. *Applied Ergonomics*, *19*(1), 55-66.

Lee, B. P. H. (2001). Mutual knowledge, background knowledge and shared beliefs: Their roles in establishing common ground. *Journal of Pragmatics*, *33*(1), 21-44.

Leong, T., Howard, S., & Vetere, F. (2008). Take a chance on me: Using randomness for the design of digital devices. *interactions*, *15*(3), 16-19.

Lutters, W. G., Ackerman, M. S., & Zhou, X. (2007). Group information management.

In W. Jones & J. Teevan (Eds.), *Personal information management.* University of Washington Press.

Magee, L. (1990). R^2 measures based on Wald and Likelihood Ratio joint significance tests. *The American Statistician*, *44*(3), 250-253.

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

Marchionini, G. (1997). Foundations for personal information infrastructures: Information-seeking knowledge, skills, and attitudes. In *Information seeking in electronic environments.* Cambridge University Press.

Mark, G., & Prinz, W. (1997). What happened to our document in the shared workspace? The need for groupware conventions. In *IFIP TC13 International Conference on Human-Computer Interaction* (p. 413-420).

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, *127*(4), 331-354.

Markus, M. L. (2001). Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems*, *18*(1), 57 - 93.

Marlow, C., Naaman, M., boyd, d., & Davis, M. (2006). Position paper, tagging, taxonomy, flickr, article, toread. In *WWW '06 Collaborative Web Tagging Workshop.* Edinburgh, Scotland.

McMillan, S. J., Hoy, M. G., Kim, J., & McMahan, C. (2008). A multifaceted tool for a complex phenomenon: Coding web-based interactivity as technologies for interaction evolve. *Journal of Computer-Mediated Communication*, *13*(4), 794-826.

McNemar, Q. (1946, Jul). Opinion-attitude methodology. *Psychol. Bulletin*, *43*(4), 289-374.

Mehlenbacher, B., Duffy, T. M., & Palmer, J. (1989). Finding information on a menu: Linking menu organization to the user's goals. *Human-Computer Interaction*, *4*(3), 231-251.

Menard, S. (2002). *Applied logistic regression analysis*. Sage University Press.

Mervis, C., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89-115.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201-213.

Miles, M. B., & Huberman, M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd Edition ed.). Sage Publications, Inc.

Mobrand, K. A., & Spyridakis, J. H. (2007). Explicitness of local navigational links: comprehension, perceptions of use, and browsing behavior. *Journal of Information Science*, *33*(1), 41-61.

Murphy, G. L., & Lassaline, M. E. (1997). Hierarchical structure in concepts and the basic level of categorization. In K. Lamberts & D. Shanks (Eds.), *Knowledge, Concepts, and Categories (Studies in Cognition)* (p. 93-131). Hove, East Sussex, UK: Psychology Press.

Nagelkerke, N. (78). A note on a general definition of the coefficient of determination. *Biometrika*, *3*, 691-2.

Nickerson, R. S. (1999). How we know – and sometimes misjudge – what others know: imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737-759.

Nickerson, R. S. (2001). The projective way of knowing: A useful heuristic that sometimes misleads. *Current Directions in Psychological Science*, *10*(5), 168-172.

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, *21*(4), 337-360.

Oakes, W. (1972, Oct). External validity and the use of real people as subjects. *American Psychologist*, *27*(10), 959-962.

Olson, G., & Olson, J. (2000). Distance matters. *Human-Computer Interaction*, *15*, 139-178.

Papadopoulos, P., Bruno, G., & Katz, M. (2008). Beyond Beowulf clusters. *ACM Queue*, *5*(3), 36-43.

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science*, *29*(3), 343-373.

Polson, P. G., & Lewis, C. H. (1990). Theory-based design for easily learned interfaces. *Human-Computer Interaction*, *5*(2&3), 191-220.

Poole, D. (2008). Breaking the major release habit. *ACM Queue*, *4*(8), 46-51.

Popovich, P. M., Gullekson, N., Morris, S., & Morse, B. (2008). Comparing attitudes towards computer usage by undergraduates from 1986 to 2005. *Computers in Human Behavior*, *24*(3), 986-992.

R Development Core Team. (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0)

Rader, E. (2009). Yours, mine and (not) ours: Social influences on group information repositories. In *CHI '09: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 2095-2098). New York, NY, USA: ACM.

Rafferty, P. (2001). The representation of knowledge in library classification schemes.

Knowledge Organization, *28*(4), 180-191.

Richardson, C. (2008). Untangling enterprise java. *ACM Queue*, *4*(5), 36-44.

Ross, N., & Medin, D. L. (2005). Ethnography and experiments: Cultural models and expertise effects elicited with experimental research techniques. *Field Methods*, *17*(2), 131-149.

Russell, A. W., & Schober, M. F. (1999). How beliefs about a partner's goals affect referring in goal-discrepant conversations. *Discourse Processes*, *27*(1), 1-33.

Russell, D. M., Slaney, M., Qu, Y., & Houston, M. (2006). Being literate with large document collections: Observational studies and cost structure tradeoffs. In *HICSS '06*.

Sanderson, J. (2008). The blog is serving its purpose: Self-presentation strategies on 38pitches.com. *Journal of Computer-Mediated Communication*, *13*(4), 912-936.

Schober, M. F., & Brennan, S. E. (2003). Processes of interactive spoken discourse: The role of the partner. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (p. 123-164). Lawrence Erlbaum: Hillsdale, NJ.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211-232.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., et al. (2006). tagging, communities, vocabulary, evolution. In *Cscw '06* (p. 181-190).

Shami, N. S., Ehrlich, K., Gay, G., & Hancock, J. T. (2009). Making sense of strangers' expertise from signals in digital artifacts. In *CHI '09: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 69-78). New York, NY, USA: ACM.

Steinman, M. J. (2008). Unified communications with SIP. *ACM Queue*, *5*(2), 50-55.

Suchman, L. (1994). Do categories have politics? The language/action perspective reconsidered. *Computer Supported Cooperative Work*, *2*(3), 177-190.

Tang, J. C., Drews, C., Smith, M., Wu, F., Sue, A., & Lau, T. (2007). Exploring patterns of social commonality among file directories at work. In *CHI '07: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 951–960). New York, NY, USA: ACM.

Tao, Y.-H. (2008). Information system professionals' knowledge and application gaps toward web design guidelines. *Computers in Human Behavior*, *24*(3), 956-968.

Tashakkori, A., & Creswell, J. W. (2007). Exploring the nature of research questions in mixed methods research. *Journal of Mixed Methods Research*, *1*(3), 207-211.

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (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 (p. 415-422). Vienna, Austria: ACM Press.

Trigg, R. H., Blomberg, J., & Suchman, L. (1999). Moving document collections online: The evolution of a shared repository. In *ECSCW '99: Proceedings of the Sixth European Conference on Computer-Supported Cooperative Work.* Copenhagen, Denmark.

Šauperl, A. (2004). Catalogers' common ground and shared knowledge. *Journal of the American Society for Information Science and Technology*, *55*(1), 55-63.

Vaughan, M. W., & Dillon, A. (2006, June). Why structure and genre matter for users of digital information: A longitudinal experiment with readers of a web-based newspaper. *International Journal of Human-Computer Studies*, *64*(6), 502-526.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217-221.

Voida, S., Edwards, W. K., Newman, M. W., Grinter, R. E., & Ducheneaut, N. (2006). Share and share alike: exploring the user interface affordances of file sharing. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* New York, NY, USA: ACM.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-196.

West, R. (2008). The psychology of security. *Communications of the ACM*, *51*(4), 34-40.

Whalen, T., Toms, E. G., & Blustein, J. (2008). Information displays for managing shared files. In *CHiMiT '08: Proceedings of the 2nd ACM Symposium on Computer Human Interaction for Management of Information Technology* (pp. 1–10). New York, NY, USA: ACM.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Neilsen & R. L. Mack (Eds.), *Usability inspection methods* (p. 105-140). John Wiley.

Whittaker, S., & Hirschberg, J. (2001). The character, value, and management of personal paper archives. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *8*(2), 150 - 170.

Whittaker, S., & Sidner, C. (1996). Email overload: exploring personal information management of email. In *CHI '96: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (p. 276-283). New York, NY, USA: ACM.

Williams, D., Yee, N., & Caplan, S. E. (2008). Who plays, how much, and why? Debunking the stereotypical gamer profile. *Journal of Computer-Mediated Communication*, *13*(4), 993-1018.

Wittwer, J., Nückles, M., & Renkl, A. (2008). Is underestimation less detrimental than overestimation? The impact of experts' beliefs about a layperson's knowledge on learning and question asking. *Instructional Science*, *36*(1), 27-52.

Workman, M., Bommer, W. H., & Straub, D. (2008). Security lapses and the omission of information security measures: A threat control model and empirical test. *Computers in Human Behavior*, *24*(6), 2799-2816.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441-517.