

UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Academic Knowledge Transfer in Social Networks

Permalink

<https://escholarship.org/uc/item/4n3332xn>

Author

Slater, Mark David

Publication Date

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ACADEMIC KNOWLEDGE TRANSFER IN SOCIAL NETWORKS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Mark David Slater

March 2013

The Dissertation of Mark David Slater
is approved:

Professor E. James Whitehead, Chair

Professor Scott Brandt

Professor Noah Wardrip-Fruin

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Copyright © by
Mark David Slater
2013

Table of Contents

List of Figures	vii
List of Tables	ix
Abstract	x
Acknowledgments	xii
1 Introduction	1
1.1 Motivations	1
1.2 Knowledge Transfer at Large	4
1.2.1 Sociology and Psychology	5
1.2.2 Media Communications	12
1.2.3 Computer Supported Cooperative Work	13
1.3 Knowledge Transfer in Digital Libraries	17
1.4 Knowledge Transfer in Social Networks	19
1.5 Research Topics	20
2 A Model of Academic Knowledge Transfer	22
2.1 Overview	23
2.1.1 Discussions	24
2.1.2 Speeches and Presentations	25
2.1.3 Electronic Messages	25
2.1.4 Written Documents	26
2.1.5 Journals	27
2.1.6 Conferences	27
2.2 Knowledge Transfer Paths	28
2.3 Discussion	30
3 A Software System for Academic Knowledge Transfer	33
3.1 Major Features	34
3.1.1 Digital Libraries	35
3.1.2 Versioned File Spaces	35

3.1.3	Projects	36
3.1.4	Community Discussions	36
3.1.5	Community Journal	37
3.1.6	Features in Relation to the Model of Knowledge Transfer	38
3.2	User Considerations and Cross Cutting Concerns	40
3.2.1	Roles and Privacy	41
3.2.2	Solo Individuals	41
3.2.3	Mobile Individuals	42
3.2.4	Interconnected Individuals	44
3.2.5	Extensible Implementation	45
3.2.6	Other Features	46
3.3	Applications of the Model of Academic Knowledge Transfer	47
4	Whisper: A Prototype Personal Digital Library for an Academic Knowledge Transfer System	49
4.1	Introduction	50
4.2	Related Work	51
4.3	Digital Library Requirements	54
4.4	System Architecture	56
4.4.1	Resource Tier	56
4.4.2	Data Management and Services Tiers	58
4.4.3	Client Tier	59
4.5	Implementation	59
4.5.1	Data Model Implementation	60
4.5.2	Controller Implementation	61
4.5.3	View Implementation	63
4.6	Lessons Learned	64
4.6.1	Implementation Choices	64
4.6.2	User Feedback	65
4.7	Interface Alterations	66
4.8	Conclusion	69
5	A Cautionary Tale of Turning Work Into Play	74
5.1	Introduction	74
5.2	Study Applications	77
5.2.1	Tidbitz	80
5.2.2	Nuggetz	83
5.2.3	Read All About It	84
5.3	Experimental Approach	88
5.3.1	Surveys and Logs	88
5.3.2	Application Users	89
5.4	Survey Results	91
5.4.1	Information Sharing Habits	91
5.4.2	Information Sharing Applications	94

5.4.3	Social Context	96
5.5	Discussion	98
5.5.1	Information Sharing Habits	98
5.5.2	Social Groups and Problems Faced	99
5.5.3	Framing Facebook	101
5.5.4	Application Design	102
5.6	Conclusion	102
6	Information Sharing on Twitter	104
6.1	Introduction	105
6.2	Related Work	109
6.3	Experimental Design	114
6.3.1	Twitter User Selection	114
6.3.2	Data Harvesting	117
6.3.3	Tweets and Retweets and Retweeted Retweets and Links (Oh my!)	120
6.3.4	Statistical Analysis	122
6.4	Results	132
6.4.1	Tweet Behavior of the users	132
6.4.2	Retweet Behavior of the followers	139
6.4.3	Tweet Content	148
6.4.4	Tweet Timing	150
6.4.5	Tweet Rate and Followers	160
6.5	Discussion	166
6.5.1	Knowledge Transfer via Tweets	166
6.5.2	Tweet Content	172
6.5.3	Tweet Timing	173
6.5.4	Threats to Validity	174
6.6	Implications for an Academic Knowledge Transfer System	175
6.7	Conclusion	175
7	Conclusion	177
7.1	Contributions	177
7.1.1	A theoretical model of information sharing	178
7.1.2	Goals and requirements for a knowledge sharing software system	178
7.1.3	A prototype software system	179
7.1.4	Gamification of academic knowledge sharing	180
7.1.5	Information sharing in Twitter and its implications	180
7.2	Future Work	181
7.2.1	Full Scale Implementation	182
7.2.2	Scientometrics and Whisper	183
7.2.3	Scientometrics and Twitter	184
7.3	Final Thoughts	185

A Facebook Survey Questions	187
A.1 Intake Survey Questions	187
A.2 Final Survey Questions	189
B List of Celebrity Twitter Users	191
Bibliography	193

List of Figures

2.1	Knowledge Transfer Mechanisms	24
2.2	Example Knowledge Transfer Paths	28
3.1	Interaction with Repositories	38
3.2	Interaction with Committees	39
3.3	Publication to Community	39
4.1	Conceptual Architecture	57
4.2	Screenshot from the working prototype	60
4.3	A sample page from the digital library test bed.	66
4.4	The redesigned digital library screen.	70
4.5	The redesigned item details page.	70
4.6	The redesigned item details page showing the open “My Work” panel.	71
4.7	A prototype item details page.	71
4.8	A prototype details page showing the open “My Work” panel.	72
4.9	A prototype edit page.	72
5.1	Tidbitz Home	78
5.2	Tidbitz Search Filter Dialog	79
5.3	Tidbitz Read Message Dialog	80
5.4	Tidbitz New Tidbit Page	81
5.5	Nuggetz Home	83
5.6	Nuggetz Ratings	84
5.7	Read All About It Home	85
5.8	Read All About It Message Dialog	86
5.9	User Institutions	90
5.10	User Genders	90
5.11	Information Sharing Method Use	94
5.12	Information Sharing Method Satisfaction	96

6.1	Tweet Distributions for each category of user. Each user is represented by a bar along the x-axis, sorted by Percent Tweets With Links. The space between the top of each bar and 100% represents the user's plain tweets.	133
6.2	Percent Tweets With Links Boxplot	135
6.3	Percent Retweets Boxplot	136
6.4	Percent Retweets With Links Boxplot	138
6.5	Retweeted Tweet Distributions for each category of user. Each user is represented by a bar along the x-axis, sorted by Retweeted Tweets With Links. The empty space above each bar to 100% represents the user's tweets that were not retweeted.	140
6.6	Percent Retweeted Tweets Overall Boxplot	142
6.7	Percent Retweeted Plain Tweets Boxplot	144
6.8	Percent Retweeted Tweets With Links Boxplot	145
6.9	Percent Retweeted Retweets Boxplot	147
6.10	Percent Retweeted Retweets With Links Boxplot	149
6.11	Tweet Content Categories	151
6.12	Hour of Day (GMT): All Users – The number of tweets posted during a particular hour of the day (GMT) for celebrities and regular users. The data points are non-continuous.	153
6.13	Hour of Day (GMT): Celebrities – The number of tweets posted during a particular hour of the day (GMT) for celebrities. The data points are non-continuous.	154
6.14	Hour of Day (GMT): Regular Users – The number of tweets posted during a particular hour of the day (GMT) for regular users. The data points are non-continuous.	155
6.15	Day of Week: All Users – The number of tweets posted on a particular day of the week (GMT) for celebrities and regular users. The data points are non-continuous.	156
6.16	Day of Week: Celebrities – The number of tweets posted on a particular day of the week (GMT) for celebrities only. The data points are non-continuous.	157
6.17	Day of Week: Regular Users – The number of tweets posted on a particular day of the week (GMT) for regular users only. The data points are non-continuous.	158
6.18	Tweet Rate: Tweets per Day	161
6.19	Tweet Rate Regression	162
6.20	Follower Count	163
6.21	Number of Followers (all users)	164
6.22	Number of Followers (regular users only)	165
6.23	A model of knowledge transfer via tweets	170

List of Tables

4.1	Library Item Metadata	55
4.2	Author Metadata	55
6.1	Regular User Selection Criteria And Rejection Counts	116
6.2	Twitter User Data	118
6.3	Twitter Tweet Data	119
6.4	Some fake Tweet data	126
6.5	User behavior in fake Tweet data	127
6.6	Follower behavior in fake Tweet data	127
6.7	Tweet Content Categories	130
6.8	Shapiro-Wilk Analysis of Tweet Types	134
6.9	Shapiro-Wilk Analysis of Retweeted Tweet Types	141
6.10	Computed Coefficients of Retweetability	168
6.11	Expected celebrity tweet audience	169
6.12	Expected regular user tweet audience	169
B.1	Technology	191
B.2	Politics	191
B.3	Business	192
B.4	Entertainment	192
B.5	Higher Education	192

Abstract

Academic Knowledge Transfer in Social Networks

by

Mark David Slater

The rise of online social networks has presented many new opportunities and methods for human communication in general and academic communication in particular. Knowledge is created at every level of academia, including individuals, project groups, and research communities. For knowledge to have lasting utility, it must be transferred from one mind to another, for only when knowledge is instantiated in someone's mind can it be used as the base for future action and thought. Today, despite decades of research, computer support for knowledge workers in academia is fragmented and poorly integrated. While office automation and other forms of CSCW approaches have benefitted academics, no single environment today integrates the basic research activities of the academic knowledge worker, including: individually or collaboratively writing research papers, sharing research papers, reviewing papers for publication in a journal or conference and persistently sharing comments and observations on existing literature.

This thesis explores academic knowledge transfer within social networks in several ways. It first presents a model of academic knowledge transfer, along with the requirements for a software system that instantiates that model. The proof-of-concept, named Whisper, for that software system is described, along with the feedback from its users and the changes made to the system's design based on that feedback.

Additionally, an experiment in gamification of knowledge transfer within the Facebook social network is explored, with implications for the knowledge transfer system design. Finally, data from celebrity users and regular users of the Twitter social network is contrasted, providing insight into how often information from the different types of users is re-shared to others, how the “packaging” of that information (a simple statement vs a link to a website) affects the re-sharing rate, and methods users might try to increase the depth their messages can reach in the Twitter network.

Academics currently have very poor tool support for some of the most common tasks they perform, such as organizing the files (both data and research output) across software applications, and linking research output back to the raw data. Both of these concerns, and others, would be addressed by a knowledge sharing environment based on the model described here. The full development of such a system presents additional opportunities for research in human factors, CSCW, and social psychology.

Acknowledgments

I would first like to thank my advisor, Jim Whitehead, for his patience and assistance over the years. This thesis was an odyssey for me, and without his support I wouldn't have stuck with it. I also need to thank my committee members, Professors Scott Brandt and Noah Wardrip-Fruin, for their advice, especially in the final stages of the research. Tracie Tucker masterfully marshaled me through the department's and university's processes, and I very appreciative of her assistance.

Over the years, I have occasionally needed assistance in the various tasks I've undertaken during this research. I would like to thank Kate Treichler and Anastassia Drofa for their assistance with the Human-Computer Interaction aspects of this work, as well as Greg Bruins and April Smith for their assistance with the icons and images used in the applications. Andrew Pilecki was instrumental in the design of the surveys within the Facebook applications. I would also like to thank Professor Abel Rodriguez, Alison Davis-Rabosky, Jen Maresh, Lesley Lancaster, and Jason Farris for their support in analyzing the data collected from Twitter.

My employer, Silver Spring Networks, kindly allowed me the flexibility in my work schedule to be able to continue my research. I have to thank my supervisors over the years for understanding that I served two masters: Don Reeves, Josh Powell, Russ Wright, Josh Atir, Dave Kong, and especially Leena Janardanan who had to deal with the final throes of finishing this work.

I have been honored with the friendship of Brianne Hunter, Sarah Thom, and David Rosario who not only supported my efforts, but also occasionally allowed them-

selves to be used as guinea pigs, software testers, and sounding boards. Finally, I need to thank my family – Sandra, David, Stephen, Andrew, Jessica, and Jude – for their patience and understanding when I had to work through family events, leave early, or miss them altogether in the pursuit of this thesis.

Chapter 1

Introduction

The rise of online social networks has presented many new opportunities and methods for human communication in general and academic communication in particular. Gone are the days (only a few decades ago) of collaboration via postal service. The speed at which knowledge flows through these new networks can be dizzying, with an idea or fact being posted by a single person and available to millions of people, both known and unknown to the original poster, within seconds. However, the application of social networks to problems that arise in academic communication is still being explored.

1.1 Motivations

Knowledge is created at every level of academia, including individuals, project groups, and research communities. For knowledge to have lasting utility, it must be transferred from one mind to another, for only when knowledge is instantiated in someone's mind can it be used as the base for future action and thought. Today, despite

decades of research, computer support for knowledge workers in academia is fragmented and poorly integrated. While office automation and other forms of CSCW approaches have benefitted academics, no single environment today integrates the basic research activities of the academic knowledge worker, including: individually or collaboratively writing research papers, sharing research papers, reviewing papers for publication in a journal or conference and persistently sharing comments and observations on existing literature.

Consider a typical collaborative conference paper. It begins as a file that is exchanged multiple times via email as each author makes contributions. The paper is then uploaded to a web-based review system. Upon acceptance, the connection between the reviews, the paper source, and the final paper version is broken, since the camera-ready paper is submitted to the publisher's own web-based paper manager. The final paper is then sent to an institutional digital library. Researchers download papers from the library, storing them in a folder on their local disk, where they are difficult to search, and are completely dissociated from their bibliographic metadata. The paper has now crossed five system boundaries, effectively eliminating any possibility of advanced collaboration such as shared annotations, and shared collection management.

Ideally, knowledge flow among people is frictionless, akin to drops of water on a teflon pan, where ideas are transferred among people quickly, with minimal effort. In reality, there is friction in the flow of knowledge, causing knowledge flow to be slower and more laborious, sometimes not occurring at all. Most academic knowledge workers make use of the same toolset: e-mail client, web browser, word processor, spreadsheet,

and presentation authoring. It is not our purpose to replace or recreate these tools, but rather to give academics a standard method for collaboratively creating knowledge that cuts across tool vendors and operating systems; once knowledge has been created or captured with these tools, we seek to improve the transfer of this knowledge to its intended audience.

The friction generated during knowledge transfer can only be reduced once the sources of that friction have been identified. When two people collaborate on a paper using email to exchange successive versions, friction takes the form of lost changes and the inability to safely work in parallel. Completed papers are frequently held to a page limit for publication; friction here results in less detail, and possibly little to no information about basic assumptions or failed experiments along the way. The business model of most academic journal publishers limits the number of papers that can be published in any one volume, and friction appears in the delay between a paper's acceptance and its publication, sometimes lasting years. If a new researcher joins a project in mid-stream, friction is the time lost as one or more existing team members acclimate the new member. Likewise, when starting work in a new discipline, friction is the time spent identifying the most salient works in that field and understanding how they fit into the context of the current research. Friction is not limited to knowledge transfer between people; it can also arise in the time and mistakes made when synchronizing working documents between computers, and materials forgotten on a computer inaccessible from the researcher's current location.

The most complex problems facing science and society today are more likely to

be solved at the intersection of two or more disciplines rather than any one discipline; a recent survey found that interdisciplinary research has increased over the last few decades [134]. In recent years, the depth of knowledge needed to succeed in a chosen field has required practitioners to become more specialized, leading to fragmentation. This fragmentation leads to multiple related disciplines that are less integrated while making it more difficult to pursue cross-cutting research as required knowledge is spread among tens of journals and conferences. Even the best search engines cannot reliably gauge the quality of the documents they return; flawed documents may be cited more often than high-quality research as members of the field rush to correct the errors. Without tools designed to address knowledge flow within and between disciplines, academics will find it increasingly difficult to function effectively in their field; the challenge for knowledge workers should be the creation of new knowledge, not keeping abreast of recent developments.

1.2 Knowledge Transfer at Large

Knowledge transfer has been studied by researchers in a variety of fields. Researchers in Sociology and Social Psychology have been examining knowledge transfer and its effect on social groups and organizations for decades. With the advent of radio and television, advertisers, politicians, and academics wanted to better understand how knowledge and opinions spread, giving rise to the modern field of Media Communications. More recently, the ubiquity of computers and their subsequent interconnection via the internet has led to the study of Computer Supported Cooperative Work (CSCW).

1.2.1 Sociology and Psychology

In Sociology and Psychology, several theoretical models were developed to describe and understand the behaviors of people engaged in knowledge transfer activities. Among the most popular of these models are social capital and networks [21, 77, 93], the theory of reasoned action [66], and social exchange theory [51].

1.2.1.1 Social Capital and Networks

Social capital theory attempts to describe the ties that bind a community together, and how people within those communities interact. From a simplistic perspective, social capital can be seen as a form of exchange in which community members help others not out of pure altruism, but rather because they had been helped in the past and might need help again in the future. Of particular relevance to this work is Granovetter’s description of “weak ties” between community members [77] and the surprising utility they provide compared to “strong ties”. The relative strength of a tie can be determined by examining the how long the tie has existed, the emotional intensity, the intimacy, and reciprocated services between a pair of people in the community. Granovetter noted that that given three people: A, B, and C, if there is a strong tie between A and B, and another strong tie between A and C, there is almost always a weak tie between B and C, and that such weak ties are “indispensable to individuals’ opportunities and to their integration into communities”. Paradoxically, he found that strong ties, while they assisted local cohesion, led to fragmentation within the overall community.

More recently, Nahapiet and Ghoshal [129] argued that social capital “facili-

tates the creation of new intellectual capital” or knowledge. This is part of their larger argument that large organizations (generally businesses), with their sense of community, can develop high levels of social capital more efficiently than the more amorphous markets the organizations inhabit. They also suggest that some organizations perform better than others specifically because of “their ability to create and exploit social capital”, especially when they specifically invest in resources to do so. Their discussion of the aspects of social capital was later built upon by Widn-Wulff and Ginman [168] who explored in more detail how those aspects of social capital affect knowledge sharing within organizations. Widn-Wulff and Ginman described the structural, content, and relational components of social capital, and identified different aspects of the components. For example, the structural component includes the social network, while the content component includes actual communication and information exchange, and the relational component is concerned with obligations and trust. They then identified possible methods of measuring these components that had previously been described in the literature, but without the benefit of the social capital framework.

Brown and Duguid [28, 29] also recognized the critical role that knowledge, and specifically its organization within the community, plays. They note that “socially embedded knowledge that ‘sticks,’ because it is deeply rooted” within the organization and its efforts, it is the knowledge that the organization is built upon, and that when the various parts of an organization lose their coherence, they are no longer able to use or transfer this knowledge effectively. By contrast, “leaky knowledge” will not flow as easily between parts of an organization that do not share a community of practice. This

type of knowledge is often more specific to the type of work being done by a unit of the organization (engineering vs. marketing vs. manufacturing, and so on) and will more easily spread to practitioners within the same field [27]; perhaps ironically, this is also the type of knowledge that organizations often see as giving them a competitive advantage and try their hardest to protect. Hansen later found that weak ties between units of an organization do promote the fast spread of simple knowledge, but impede the fast spread of complex knowledge, while strong ties impede the spread of simple knowledge but speed the spread of complex knowledge [79]. Complex knowledge is often leaky knowledge because it is more easily transferred to people with strong ties due to similar training or experience than it is to those without the background to understand it. Hansen later extended this work to include indirect relationships via the social network within the organization and found that shorter network paths between units assisted in knowledge transfer [80].

In their study of a popular web message and discussion forum for legal practitioners, Wasko and Faraj [165] explored the motivations people had for sharing knowledge within their community of practice. They found that benefits to a individual's reputation was a significant motivating factor. Other important predictors of contribution included an individual's experience in the field, and the people with large connections throughout the community. They also found that high levels of the relational component of social capital did not predict contribution, and attributed this finding to the electronic and impersonal nature of the forum.

Wu et al. used social capital to explore how managers could encourage knowl-

edge transfer [172]. They emphasized affect-based trust and social interaction as the primary components that form the structure of social capital that foster the determinants of knowledge transfer, knowledge sharing and a new variable they call learning intensity, which can be described as people’s motivation and ability to learn the knowledge being shared. In a field study, they found that higher levels of affect-based trust lead to both greater knowledge sharing and learning intensity. Additionally, social interaction had a positive affect on learning intensity, but did not affect the level of knowledge sharing; however they note that “social interaction is more concrete, more controllable, and easier to implement” than trust.

1.2.1.2 Theory of Reasoned Action

The theory of reasoned action (TRA), developed by Fishbein and Ajzen [66], contends that a person is very likely to do things they have the behavioral intent of doing. Behavioral intent is formed by the person’s attitude about that action and the subjective norms inherent in their community. This means, for example, that a person must have a very strong positive attitude about an action to form the intent to do so if the community in which they are acting views that action in a negative light; in simplistic terms, people generally won’t do things that they’d be judged negatively for. Likewise, if their community views an action in a particularly strong light, a subject is likely to form the intent to participate in that activity unless their attitude is just as strongly negative about that activity; this is one way of describing peer pressure.

Bock et al. [20] based their study of factors that inhibit or strengthen a subject’s

intention on TRA, arguing that unlike fields in which the behaviors and actions are fully explored, when the action is knowledge sharing the factors are not yet well understood. Based on interviews with executives of several large organizations, they found both external and internal motivators which were incorporated into a model which extended the original TRA with the addition of an overall organizational climate, subsequently tested in a field survey. They found that anticipated reciprocal relationships, an internal motivator that fed a subject's attitude towards knowledge sharing, was the primary driver supporting the action. An increase in the subject's self worth, another internal motivator, was determined to be a significant factor in the subjective norms of the community. Additionally the organizational climate, which included external motivators such as the overall fairness of, affiliation with, and innovation within the organization, had a very strong influence on the subjective norm, as well as influencing the subject's eventual intention to share knowledge (though to a lesser extent). External economic rewards were found to have a negative influence on a subject's intent to share. Their study was perhaps the first to demonstrate that "the institutional structures within which a focal behavior is situated also influence behavioral intentions."

A study performed by Hsu and Lin [86] about the acceptance of blogging is of particular interest here because of the close parallels between participating in the blogging community (as a reader or as a writer) and participating the Twitter community. Blogging on the internet doesn't have the same organizational context that knowledge sharing within business does, so the organizational climate described by Bock et al. [20] is not present in Hsu and Lin's model. In a field survey, they examined how techno-

logical aspects and knowledge sharing aspects affect an individual's attitude towards blogging. They found that enjoyment of the technology had a strong effect, as did ease of use, on a person's attitude towards blogging. In the knowledge sharing aspect, they determined that altruism and reputation were good predictors. Additionally, they found that identification with the community had a greater influence than the subjective norm described by the traditional TRA model, possibly because the internet provides access to both wider and more specific social groups than people generally had access to when the TRA model was initially described. Put another way, it would seem that a large enough set of weak links formed by the internet's blogging community can have a greater influence on a person's intent to blog than the strong links that make up their subjective norms.

1.2.1.3 Social Exchange Theory

Social exchange theory "might be described, for simplicity, as the economic analysis of noneconomic social situations." [51] Just like economics, the worth of an activity can be computed based on the rewards minus the costs. In social settings, the rewards might include trust, acceptance in a community, or reciprocation, while the costs would be the effort involved in that activity, whether physical, monetary, the dedication (time) required, or even negative aspects of the other people involved in the activity. Activities in which the rewards outweigh the cost are likely to occur, while those in which the costs outweigh rewards are not.

Social exchange theory has been used to explore dilemmas including the Pris-

oner’s Dilemma and the Tragedy of the Commons [105, 174], but of particular interest is Cabrera and Cabrera’s examination of knowledge sharing dilemmas [32]. In their paper, Cabrera and Cabrera cast knowledge sharing to the public goods dilemma and explore how organizations can structure themselves to ensure the best individual (and organizational) outcome from the knowledge sharing. For individuals, the costs associated with sharing knowledge with the organization at large can overwhelm their personal rewards. Some methods of overcoming this include “emphasizing the benefits associated with exchanging personal insights”, “making people aware of the impact that their engagement in information exchanges can have on the performance of others”, and fostering a “sense of group identity and personal responsibility.”

Kankanhalli et al. developed and tested a model to explain the usage of electronic knowledge repositories, identifying the costs and rewards associated with doing so [99]. They found that the costs included the effort to transcribe knowledge into the repository, as well as the potential loss of trust they could suffer should the knowledge they share be incorrect in some way, or misused by others. Like Hsu and Lin’s study of blogging [86], they found that enjoyment in helping others was a strong reward, as was the self-confidence they gained from sharing. Additionally, organizational rewards and reciprocity from others within the organization were also useful rewards, but mostly when the organization was pro-sharing.

1.2.2 Media Communications

Researchers in media communications have long been concerned with identifying speakers, listeners, methods of communication, and the efficacy of those methods. One of the early pioneers in the field, Harold Lasswell, phrased this as determining “who says what to whom in what channel with what effect” [111]. Media Communication theories of the last century were often concerned with how much influence the media actually has on the public at large. The concept of “direct effects” suggests that the media has a very strong influence, and that people are, in some ways, more affected by the media than by those around them [91, 176]. A counter theory suggests there was more subtlety and found that while the media influences the public, it is assisted by local “opinion leaders” who act as intermediaries or amplifiers for the media’s message, making the information flow a two-step process [100, 113]. Recent research gives more credence to the two-step (or even a multi-step) process, because the rise of the internet, and especially social networks such as Facebook and Twitter [10], enables and encourages conversations that “guide media consumption simultaneously” [162]. Twitter, along with other microblogging tools, represents a new channel of communication in which theories of mass communication can be explored.

Craig published a review of communication theory research in which he attempted to unify the traditions of the field into a cohesive model [42], noting publications in the field “seldom mention other works on communication theory except within narrow (inter) disciplinary specialties and schools of thought” due, at least in part, to the many disciplines which have contributed to the field’s development. He identified

seven traditions of communication theory: Rhetorical, Semiotic, Phenomenological, Cybernetic, Sociopsychological, Sociocultural, and Critical. The Rhetorical tradition (as one might expect) is focused on the art of speaking or writing, often with the intent to educate or persuade. One might view Twitter as a platform for “soundbites” in this view of communication, while YouTube and blogs provide a platform for long-form communication; in both cases the way in which the communicator expresses themselves, and the reaction evoked from their audience, can be assessed. Of particular import is the Sociopsychological tradition, in which “communication ... is the process by which individuals interact and influence each other” and generally examines expression, interaction, and influence; social media is useful in this tradition because it allows those interactions, as well as the ebb and flow of influence within the system, to be studied in detail.

1.2.3 Computer Supported Cooperative Work

Computer Supported Cooperative Work (CSCW) is the field of Computer Science and Computer Engineering which focuses on, among other things, enabling knowledge sharing in various environments (work, school, the internet, etc.). CSCW research is not so much concerned with *why* people share knowledge in general, as it is with *how* they do it, specifically within communication channels created or enabled by computer hardware, software, and networks. Some examples include:

- Modeling the sharing process and creating frameworks or tools based on those models, enabling groups to share knowledge and learn more effectively [97, 102,

163] and even the original proposal for what would become the world wide web [13].

- Direct exploration of a theory from an external field, such as the recent work from Shen et al. that applies economic theory to knowledge sharing [145] and found virtual currencies an effective means of encouraging people to share.
- Examining the output of a particular design methodology, such as the corporate groupware system developed by Miller et al. [127] using Value Sensitive Design [67]. Value Sensitive Design tries to identify features that encourage use (such as knowing how often peers used your contribution) and features that would discourage use (such as anonymous posting) during the design process.
- Studying how work processes affect social factors in a group, such as how Agile development [123] supports knowledge sharing [117] or how it can be adapted to fit different software development team structures [112].
- Developing new hardware that enhances people's ability to share knowledge in various contexts, such as museums and conferences [152] or schools [24].

However, it is the research being done around social networking, and features of knowledge sharing that is most relevant to the work presented here.

boyd and Ellison [23] provided an in-depth history of social networking in the web, including their feature sets, how they rose to popularity, and how they fell from grace. Some research has focused on the examining the connections between people on social networking sites. For example, Lampe et al. found that Facebook users are more interested in finding out more about people they know offline than developing new

connections with people online [109], and Hewitt and Forte explored how connections in Facebook affected undergraduate education and the student / faculty relationship [83].

Others have explored the use of social networking tools within the corporate environment [95, 146, 178]. Skeels and Grudin explored social network use at Microsoft, noting “the principal work-related benefit of social networking software was in the easy, unobtrusive creation, maintenance, and strengthening of easy ties among colleagues. This does enable more efficient interaction, but it has other significant benefits.” [146] For example, they found that younger professionals (those with some work experience, but without a substantial career history) get a great deal of use from LinkedIn, a professional social network built to help match job seekers with job openings, because they “anticipate being on the job market in a few years”. Older employees they spoke with, being more established in their careers, were less likely to see the benefit of having a constant supply of potential job openings to consider, while people who are just graduating have often not heard of LinkedIn or felt like they were not its target audience. Facebook and MySpace use, which was also higher among younger employees, provided more support for maintaining awareness of others within the employee’s network, both at work and outside of it. However, the broader scope of these sites, and the concerns about confidentiality created some tension for these users. This tension worked both ways, as they were concerned about accidentally disclosing internal corporate information to the outside world, and having their events (or people) in their personal lives affect how they were viewed in the workplace. These tensions were echoed in Zhao and Rosson’s study of Twitter usage in the workplace [178]; while knowing more personal

details about their co-workers helped strengthen ties between them, the public nature of Twitter precludes sharing of more detailed work-related information and thoughts.

A common theme in CSCW is to explore how particular features of knowledge sharing systems affect the process and user behavior. For example, early work from Trevor et al. [154] examined the requirements of a shared object service and proposed the use of “adapters” to help manage access to the digital artifacts within their system, allowing functionality such as dynamic adjustments to permissions and context sensitive views of artifacts.

Covi noted [41] many academics avoid digital publications both as consumers and contributors for social reasons rather than technical ones; for example, it is unclear that all universities would give work published in an electronic journal the same weight as work published in a more traditional print journal when a researcher is considered for tenure. At the time of Covi’s study, the internet was a relatively new phenomena to many people; more established academics at that time were resisting learning new technology until they could be sure it would provide a significant return.

Privacy is an important topic in both research and the media. Users of social networking sites often form protest groups (ironically, they often use the site itself to do so) when a site adjusts its interface [104], adjusts privacy controls [122], or changes the information advertisers and third-parties are allowed to access [114]. At the same time, social network data can be mined to provide automatic access controls to information in a system [73]. Razavi and Iverson noted “not everything is to be shared with everyone” [137]. They found that that privacy preferences for a given artifact can change over

time, which is challenging for users to manage, and some systems don't provide a very clear distinction between related features (such as "groups" and "communities" in their study system) which can lead to a loss of privacy control for the user (or the decision not to participate).

1.3 Knowledge Transfer in Digital Libraries

Digital libraries are generally electronic counterparts of traditional libraries; most work to date has focused on this transition [17]. Because of their lineage, digital library research isn't generally concerned with augmenting knowledge transfer beyond providing digital access to material that was traditionally distributed in print. Some have added the ability to associate personal notes with items in the library, and send links to the items as well, but most implementations of these features are rudimentary. Most digital libraries are maintained and organized by the institutions that provide them.

For example, the ACM and IEEE both provide digital libraries to their members, while businesses and organizations involved in book or journal publication frequently require payment to download the text of each article or journal issue. The Open Archives Initiative (OAI) [108] is working to develop standards for open protocols that will enable communication between various digital library installations. There has been some work to develop personal spaces [63] within the context of these institutional libraries and across them [139], but these efforts have mostly focused on customization and saved searches within the system. Sharing and personalization capabilities

within institutional libraries are extremely useful within that context, but only provide marginal assistance to the individual trying to create and maintain a digital collection of research materials.

CiteULike [36] is a digital library service focused more on organization for the individual than publication for an organization and is closest in functionality to our proposed system. Users can add content to their personal library from a large number of external digital libraries, including the ACM and IEEE digital libraries, Amazon.com, and popular research journal publishers; items not found in one of the supported sources can be added manually by the user, but are not made available to other users in site-wide searches. Users can add notes, both private and public, to their library entries, but there is no granularity here; either a note can be seen by only the user, or it can be seen by everyone. Users can also create and join groups via which they can post comments about articles and participate in discussion forums. CiteULike's focus on the individual makes it a much better tool than the personal features in the larger institutional libraries, but there are still areas for improvement.

Mendeley [126] is one of the more advanced digital libraries available today and includes social networking features. Mendeley includes web-based storage space for a personal library of PDFs, and allows users to create and join groups (public and private) in which those PDFs can be shared. Documents in a user's digital library can be accessed on any computer with the Mendeley software installed, as well as via the web and mobile operating systems. Users can add each other as contacts, and exchange private messages. They may also form groups and share documents with those groups,

as well as collaboratively annotate them, and discuss them.

1.4 Knowledge Transfer in Social Networks

The most popular social networks today include Twitter, Facebook, Google+, Digg, Slashdot, and others. They tend to be purpose built for a relatively small subset of communication activities. Twitter allows anything to be shared with others, as long as it fits into a 140 character limit. Facebook and Google+ enable people to easily stay in contact with their larger social circle, sharing events, links, news, and photos. News aggregation sites such as Digg and Slashdot allow people to highlight articles online that they find interesting and which their friends might also enjoy.

All of these sites can be used to assist knowledge transfer, but none of them are purpose built for it, especially from the standpoint of academic knowledge work, which requires a knowledge permanence and discoverability that is not generally provided by the social networks themselves. “Scientific work produces a chain of interrelated, mutually constitutive artifacts which support the proper interpretation of scientific data, and the ongoing crystallization of scientific findings.” [131] In most existing sites, articles and files are usually not stored within the network, but rather externally linked, and while the entire history of a user’s interaction with the network may be available, it is usually not searchable which makes finding older content more difficult. Even if those limitations were overcome, or proved not to be cumbersome, there are still open questions about how users within a network influence each other, and how readily they share information with their branch of the overall network. Additionally, “different scientific

practices produce quite distinct epistemic cultures, and hence the sort of knowledge that might flow readily within one culture will not flow uninhibitedly between two” [29] and there are no current tools available that actively work to improve knowledge flow across scientific cultures.

1.5 Research Topics

Given that existing social networks generally aren’t designed to support academic knowledge workers, what additional features are needed for a social networking site to explicitly support the needs of academic information sharing? To begin answering this question, a model of academic knowledge sharing is needed, with special attention paid towards aspects and methods that are supported, and not supported, in existing social networking platforms. Given this abstract model, an attempt can be made to realize it in a working software system. This was undertaken as part of the work for this thesis, but found to be beyond the scope of an individual working alone; building a complete application and a strong user base for it would take a considerable amount of effort and time. Another approach, also taken during this study, is to use existing social networks, leveraging their users to explore some implementation questions. For example, how can people be motivated to share information with people in their network with higher quality (information the recipient would find interesting or useful) and more often? Also, the differences in how information flow within existing social networks when that information comes from a media figure versus a non-media figure who is likely to have a stronger tie to the recipient.

Chapter 2 details a model of knowledge transfer within academic disciplines, while Chapter 3 presents the goals for a software system instantiated from that model, including the motivations, requirements, and the case for an integrated system, rather than tying together existing services from around the web. A software system called Whisper is presented in Chapter 4, providing a proof of concept for the Personal Digital Library aspects of the model, including the description of requirements for embedded tools for collaboration with others. Chapter 5 uses example Personal Digital Library applications built within an existing social network, Facebook. These applications each add elements taken from social games to examine how they might affect usage; unfortunately the user base was not large enough to draw any conclusions related to knowledge transfer, and possible reasons for that are explored. Chapter 6 explores the knowledge transfer that occurs in an established large scale social network, Twitter, and how a celebrity's influence within the network compares to regular users. Given these studies in existing social networks, Chapter 7 concludes with implications for both academic information sharing in general and the development of an integrated system to support it.

Chapter 2

A Model of Academic Knowledge Transfer

Due to the finite nature of software systems, even if they are extensible, they cannot be expected to capture and support every method of academic knowledge transfer that exists, or might exist in the future. By developing a model of that transfer, we are able to find and focus on the areas in which such a system can have the greatest impact. Understanding when knowledge is transferred in the process of conducting academic work, and why it is transferred at that point in the process, will enable the development of a more useful system.

There are physical and economic factors that software simply cannot directly account for; for example, a software system cannot capture knowledge transferred via hand-written notes, nor can it alter the value a tenure committee places on the various venues for publication. However, a model focused on the pathways that knowledge

takes as it is developed and published in the academic environment, as well as the artifacts used to traverse those pathways, can provide insights about the feature set of the software system. Additionally, the model enables strong support for pathways and activities specific to the academic environment that might not be included in more generalized knowledge sharing systems (frequent peer-review of knowledge artifacts, for example).

The model presented here endeavors to present generic actors, whether an individual or collective, within the academic community. The transfer paths between those actors are generalized enough to include more traditional electronic communication (email) and newer methods like micro-blogs. It should be generally useful for any software system supporting academic knowledge transfer, and provide a basis for creating additional systems in the future, as well as a framework for comparing such systems.

2.1 Overview

Individuals are the building blocks of academic entities; they may work alone or in project groups, and they frequently belong to multiple research communities. Knowledge is created and organized within all three of these entities, but to move between them, a transfer mechanism must be employed. This mechanism can be as simple as a conversation between two individuals, or as complex as a conference with thousands of attendees.

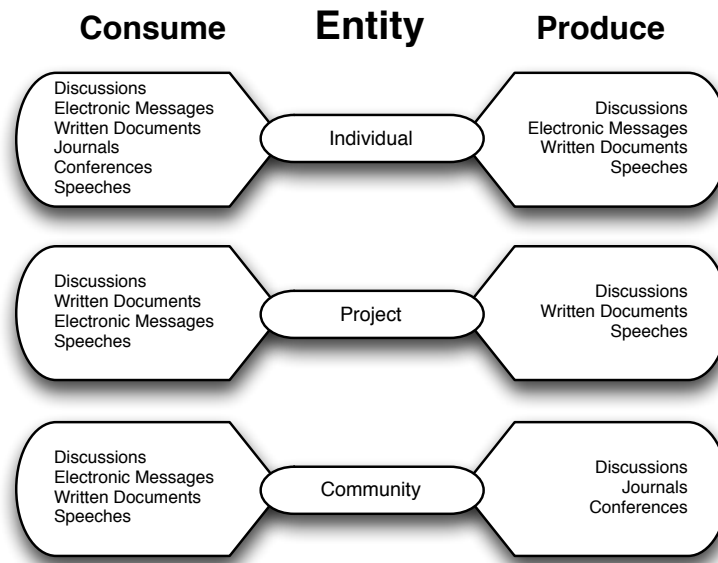


Figure 2.1: Knowledge Transfer Mechanisms

2.1.1 Discussions

Discussion is, perhaps, the most basic framework for human communication, and one of the most widely used knowledge transfer mechanisms; over time, some forms of discussion have become highly formalized, such as debates. Discussions frequently take place in real-time with all participants in close proximity. Letters, first handwritten and later typed, enable discussions when participants are geographically distant, with the added benefit of providing a persistent record; instant messaging systems are able to support the persistence of a letter while also allowing real-time interaction. The telephone provides a method for real-time discussions with a small number of geographically dispersed participants, but without a persistent record; internet voice and video conferencing systems have similar characteristics. The NLS Journal [56], bulletin board

systems, newsgroups, and blogs enable discussions to occur in near real-time, persist over time, and allow wider participation than most previous methods.

2.1.2 Speeches and Presentations

Speeches and presentations are generally given by a small group of people to a larger audience. They differ from discussions in that the information flow is mostly in one direction, from the small group to the larger. Teachers in a classroom, politicians in the houses of government, and newscasts are common forms of this type of knowledge transfer. Before Edison, the only method of recording a speech or presentation was transcription, where (most) everything that spoken is written down. The ability to make audio and visual recordings, and then copy those recordings, allows people who were not present at the time of the event to hear and see it at a later time. As internet connection speeds have increased in people's homes, more speeches and presentations are recorded and made available for replay or download.

2.1.3 Electronic Messages

E-mail is a pervasive form of electronic communication allowing people to communicate with others quickly, even at great distances. Outside non-solicited and commercial e-mail, most messages are sent to small lists of recipients (frequently only one); the sender has the benefit of knowing who the message is sent to, though there is no way to prevent a recipient from forwarding the message to others. However, it is generally considered a social faux pas to send an e-mail to large groups of people who may not be

interested in the content. For this reason, large scale e-mail messages are generally used to communicate with geographically local groups, such as the occupants of a building, rather than research communities with members in different parts of the world; as the size of a community increases, e-mail becomes a less appropriate way to communicate with members.

Whereas e-mail can be characterized as the sender “pushing” communication onto the receiver, other electronic methods allow the receiver to “pull” the communication and are considered more appropriate for large distribution. Electronic bulletin boards, dating back to the NLS Journal [56], have enabled users to post information that can persist indefinitely. Today, newsgroups allow users to post semi-persistent messages that are distributed globally, though many organizations have newsgroups that are only available to local users. Many web sites encourage users to post messages in forums related to content on the site. Blogs allow users to publish entries to websites and encourage readers to post replies and discuss each weblog entry.

2.1.4 Written Documents

Written documents, while not invulnerable, can last hundreds or thousands of years. Electronic data stores may be replacing paper as the preferred method of recording and disseminating information, but people are unlikely to stop using written documents any time soon. Paper documents do have some advantages over electronic ones. People frequently use written documents as temporary storage when moving information between electronic devices that cannot connect to each other directly. In

addition, written documents can be used in circumstances where electronic devices would fail or be impractical, such as extreme temperatures, under water, or when power for computers is unavailable.

2.1.5 Journals

An academic journal is a type of written or electronic document (sometimes both) used by communities to amalgamate and disseminate knowledge to their members. The contents of journals are generally shorter documents, each produced by a different set of authors. Most academic journals require peer-review of the submissions before a document is published to the greater community. This process helps to ensure the authors have not made any gross errors and have met a minimum standard for academic quality, as determined by the journal's editors. Journals also provide classification of and a reference point for knowledge that can be used by others when searching for related or prior work in a particular field.

2.1.6 Conferences

Conferences are similar to journals in that documents are submitted by individuals or groups for inclusion in the conference, and the submissions are generally peer-reviewed before being accepted. While conferences generally publish the accepted documents, sometimes as one edition of a related journal, they are distinct from journals in that they bring the authors of the accepted documents and other community members together in the same location for some period of time. Conferences usually provide a

forum for the authors of accepted documents to present their work to community members in person, often giving more details or highlighting portions they think will be most interesting to the greater community. Conferences also serve as social hubs for members of the community, enabling members to make new connections with people who share their academic interests. Conferences can range from small events with tens of attendees to massive productions and thousands of community members and interested outsiders.

2.2 Knowledge Transfer Paths

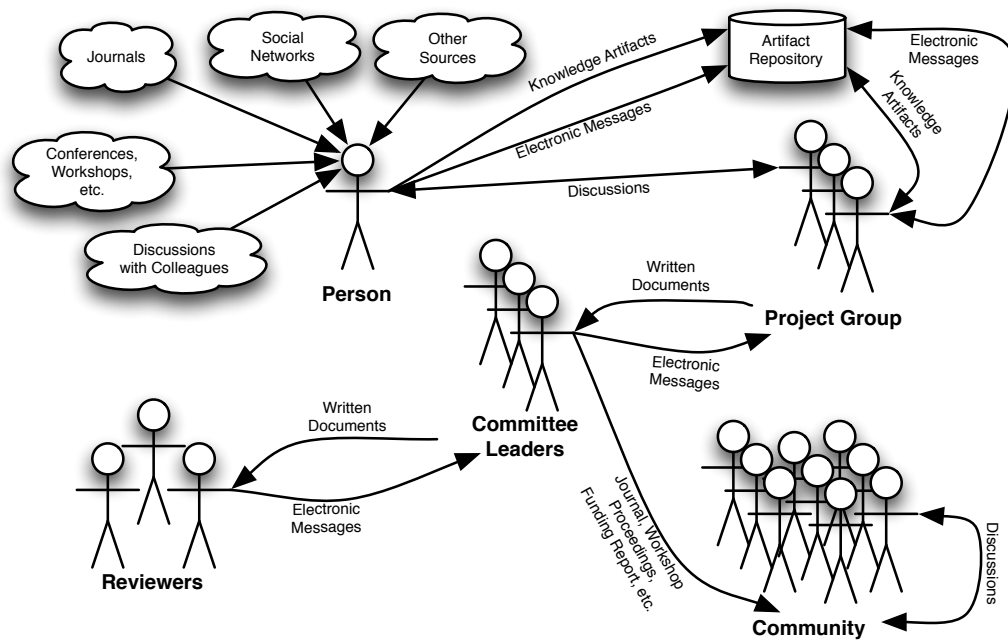


Figure 2.2: Example Knowledge Transfer Paths

Figure 2.2, shows the life-cycle of an academic paper in terms of the knowledge that goes into it and the knowledge produced by the process. One or more individu-

als draw on knowledge from any number of sources and incorporate the ideas into a project. During the project, related knowledge developed in other projects is incorporated into the new one. In addition, the participants will develop new knowledge as part of their work in the project; the diagram shows this knowledge being stored in a project repository, which can incorporate both physical and electronic storage mechanisms. Most projects eventually organize the new knowledge into one or more cohesive documents that will describe their newly developed or discovered knowledge. Sometimes these documents are shared only with individuals, as is the case for student projects in a classroom, while others are fed into, or subsumed by, other projects. The figure, however, describes the path taken to publication in a community journal or conference. In this situation, the document is passed to the group in charge of organizing the journal or conference, represented as another project group. The organizers will often send the document to individuals within the community for review, who respond with their opinions of the document (another kind of knowledge). Assuming the document is accepted for publication without additional revisions (frequently, this is not the case), the organizers of the journal or conference will publish the document, distributing it to the greater community. The case study presented here does not demonstrate the entirety of knowledge flow; there are many other paths knowledge can take as it moves from one entity to another.

2.3 Discussion

The model presented here has been developed around the specific mechanisms used to transfer knowledge between academics. Knowledge transfer studies in Sociology and Psychology (see Section 1.2.1) are generally focused on the individuals involved in the knowledge transfer, and less concerned with the mechanism of the transfer. Systems built on top of the model described here would likely provide a rich data set for exploring the relationships between users and the roles they play in knowledge transfer within the system. Social capital theorists, for example, could examine the mechanisms employed within a community, and the frequency of their use, to determine the strength of the ties between the users. A study based on the theory of reasoned action might explore the social acceptance of using a particular transfer mechanism between different research communities. Insights gained from either type of study would be very useful during the process of iteratively improving the system — building features that gently encourage additional interaction between users — but cannot inform the model until the data needed exists within a system based on the model.

A study based on social exchange theory might provide insights useful while building a system from this model, in that one of the primary goals of such a system would be to reduce the costs associated with a particular knowledge transfer activity. However, most of the costs and rewards in academics are extrinsic to the model presented here. For example, compared to sending an email, the larger costs and rewards associated with writing and publishing a journal article come from the process of passing editorial review and the increased name recognition the authors might have after publication.

The human aspects of those costs and rewards are not necessarily affected by use of a software tool; authors still have to respond to editors and reviewers, and community members still have to read the publications, regardless of how easy the physical transfer of the paper’s contents is made. Put another way, the goal of the model is to identify knowledge transfer paths that can be made smoother (less costly) through the use of a software system based upon that model. When the costs and rewards of a particular form of knowledge transfer have more to do with the people using those paths and less with the communication paths themselves, the model will be less affected by economics style analysis.

The knowledge transfer model lends itself much better to media communications examination. Based on that school of thought, one could view the model as capturing the most common channels used by academics to communication, and focusing on those channels which might be instantiated within a software system. The model supports a two-step process of media influence in which local opinion leaders might be research PIs, as well as a multi-step process that allows influence to pass through research communities and project groups. Among the traditions of communication theory identified by Craig [42], the model supports a variety of communication methods that can be assessed with Rhetorical and Sociopsychological methods.

While there is no explicit mechanism in the knowledge transfer model to support opinion leaders, the variety of supported communication paths allows interaction at multiple levels. At the local level, a user might find themselves an opinion leader if they have a personal repository in which their colleagues exchange messages, or if

they participate or lead multiple project groups. More globally, a person serving as a Committee Leader in several aspects of an academic community — funding committees, journal editorial boards, conference committees, and so on — or across multiple communities has the ability to influence a large swath of people. Through interacting at the community level, and across communities, opinion leaders within this model may be more likely to develop a multi-step process of media influence, simply because there are more opportunities for them to interact with each other.

Chapter 3

A Software System for Academic Knowledge Transfer

In 1973, Engelbart presented the idea of a Knowledge Workshop as “the place in which knowledge workers do their work” [58], and presented his concept of a computer environment that could fill this role. The internet, and more specifically the World Wide Web, presents an opportunity to create a modern Knowledge Workshop that provides users with a view of their work that is consistent between systems and builds on their existing computer interaction skills. Another benefit of the internet platform is the longevity it offers; while the life-span of a particular software system tends to be longer than the hardware upon which those systems are initially built, open protocols used to communicate on the internet tend to outlive the software systems that use these protocols.

This chapter presents a type of Knowledge Workshop based on internet- and

web-enabled software tools, with a feature set that corresponds to the model of Academic Knowledge Sharing presented in Chapter 2. Engelbart had the luxury of being one of the first to design a system of this type; he was able to design the entire workflow of the user, from their interaction with the computer to the protocols used to move information from one group’s system to another. Many of the implementation details Engelbart developed are still in use today, though they have evolved, or in some cases, devolved, significantly: networking, hypertext, graphics handling, e-mail, threaded discussions, and even the venerable mouse. But these technologies evolved separately, instead of as a cohesive whole. The end result is that today’s users have a very different, and generally more sophisticated, set of tools and skills from the users targeted by the NLS and Augment systems; they also have different (perhaps greater) expectations of the software systems they use.

3.1 Major Features

A major objective is the reintegration — in some case, the re-envisioning — of the appropriate tools for a modern Knowledge Workshop targeting academic users, resulting in a better, more efficient, work environment. A software system for this environment must enable persistent, near-realtime discussions with an individual, within project groups, and within communities. Individuals and projects need to be able to search for existing knowledge related to their work and manage those sources, along with related discussions or notes; in addition, they should be able to manage new knowledge as it is created and more easily share it with the greater community. Communities should

be allowed to explicitly examine their knowledge framework based on the publications of the community members, enabling them to better understand their history and plan future research directions.

3.1.1 Digital Libraries

A knowledge transfer system should provide users with a digital library in which they can store documents and associated commentary. These digital libraries will likely create a new first step when academics using the system begin searching for related work, and that people with a close working relationship will be more likely to search each other's libraries than those belonging to users they are unfamiliar with. For example, a professor will tend to accumulate a large digital library with commentary for many documents, both of which might be visible to their close associates, if not the entire world. The professor's students will be likely to search this digital library before looking at journals and conferences in their field because the documents stored by their professor have been vetted by a known, trusted expert, and because their professor's library will generally contain documents more related to their work than other sources.

3.1.2 Versioned File Spaces

Technically adept academics may have heard of, and even make frequent use of, versioned file spaces, but most academics either manually version their documents, or rely on their authoring environment to manage the versions. While these systems may work well enough for an individual, academic projects frequently consist of a group of

people working together. As demonstrated within the software industry, group projects are better served by an explicit version control system supported by their work environment, which in this case is the knowledge transfer system. Group members will be able to synchronize their work with each other, as well as between multiple computers, such as a desktop and a laptop; projects with only one member will also benefit from this synchronization capability. In addition, users will no longer need to rely on manual methods for differentiating versions of a document, such as including a version number in the file name.

3.1.3 Projects

Project spaces in the system should allow an individual or group of users to collect and organize their research. Projects have Digital Libraries and Versioned File Spaces distinct from those belonging to the participants of the Project; this allows a Project's lifespan to continue beyond the participation of its initial participants, and gives new project members the ability to see everything that happened before they joined the Project. Projects also have public and private Wiki-style web pages, allowing them to organize and capture details about the project for both external outreach and internal communication.

3.1.4 Community Discussions

Conferences are the most popular form of community discussion in the academic world, probably because it allows most members of the community to get away

from their normal work environment and spend time in faraway and exotic parts of the world, such as Houston, Detroit, and Fresno. However, conferences are expensive to hold and attend, so they are not frequent. While some members of a community might interact often with other members, it is difficult for an entire community to be involved in regular interactions. A Community discussion area should provide academic communities with a method of increasing the interaction between their members. While any one member might infrequently post in the discussion areas, large Communities might generate several topics of discussion each week, ranging from finding a location for the next conference to new legislation that affects research conducted by the Community. The asynchronous, open discussions provided by the system would have the effect of bringing members of a community closer together. When combined with the group project space, these spaces will foster collaboration between geographically dispersed community members by reducing the barriers to their interaction.

3.1.5 Community Journal

Each Community created within the system should be able to create one or more digital journals, allowing members a convenient means of publishing their work. Journal publications may be either complete papers or references to work published in an external Journal, whether online or print. The system will support a peer-review mechanism which the Community may use for some or all of its Journals. The Journal's Editor may choose to make the reviews of published papers public or keep them private. Each Journal publication can contain a link back to the Project within the system in

which the research was performed, thereby giving readers the ability to find additional context and details that may not have been included in the published paper.

3.1.6 Features in Relation to the Model of Knowledge Transfer

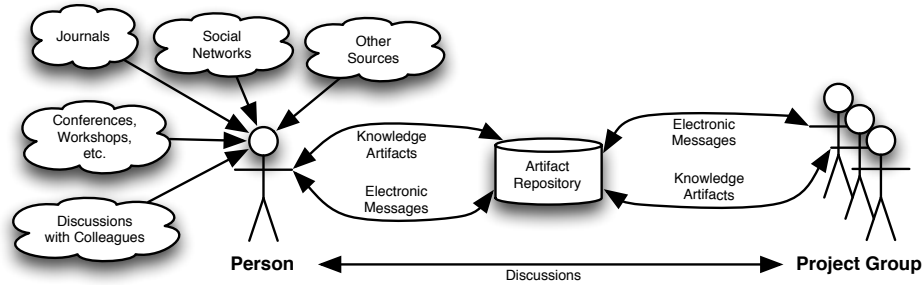


Figure 3.1: Interaction with Repositories

Individuals and Project groups both have two types of Artifact Repositories: a Digital Library (see Section 3.1.1) and a Versioned File Space (see Section 3.1.2). As an Individual finds new papers or source material, they might add them to their own Digital Library, as well as the Digital Library for any Project groups the new material is relevant to. Adding new items to the Project’s library may happen after some discussion with group members, or the addition may trigger comments (electronic messages) to be added to the new artifact within the repository. This information flow could be mirrored when an addition is made to the Versioned File Space; the major difference between the type of repository in the system is that objects in the Digital Library are expected to be static, while those in the Versioned File Space are dynamic and can be altered over time.

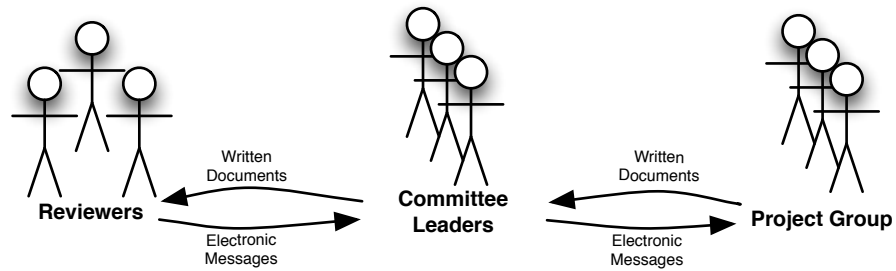


Figure 3.2: Interaction with Committees

Almost all academic work is filtered through one or more peer-review processes, generally when requesting funding for a project and when publishing research results (see Section 2.1.5). The Project group (which could be a single individual) submits research documents to an entity that solicits feedback from a set of independent Reviewers with expertise in the field. The Reviewers provide feedback to the people (or person) in charge of managing the process, who then forwards that feedback to the Project group. This process may be repeated several times until the Project’s document is accepted or finally rejected.

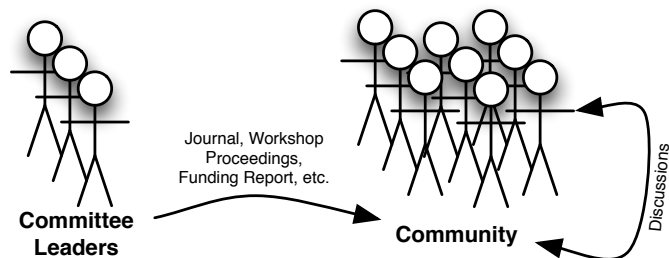


Figure 3.3: Publication to Community

When a Committee decides a document submitted by a Project for consideration is acceptable, they publish it to the Community. This publication could take the form of a simple report stating the research being funded, linking back to the Project's proposal document, or a community Journal containing papers describing new research from various Projects. The Community is then free to examine the documents and engage in discussion amongst themselves. At this point, Individuals within the Community are able to begin the cycle anew, adding the newly published document to their personal or Project repositories.

3.2 User Considerations and Cross Cutting Concerns

In determining the general requirements for this system, there are security concerns and three working environments that should be considered. In any multi-user system, controlling access to data is paramount, especially in the academic environment where research success depends on being the first one to publish. The first environment that must be considered is a user's home or office, where the bulk of their work is performed, generally as a solo task. The mobile user exists in an environment that spans multiple locations, and often relies on shared computing resources or mobile devices. Finally, there are additional needs for the interconnected user who is part of one or more larger organizations that provide the context for their work.

3.2.1 Roles and Privacy

Not all research, either in progress or finished, is intended for publication. Sometimes, it might be for use only within the organization that performed it, and other times it could be classified by a sponsoring government. More commonly, research is often not made public until it is complete simply to avoid being scooped by other researchers. Regardless of the underlying reason, privacy and security is an important aspect of the research process. Users can be assigned Roles which in turn determines the access they have. For example, a user with the “Journal Editor” role for a Community’s journal has access to submissions that have not yet been published and the peer-reviews of those submissions. Likewise, a user with the “Project Owner” role for a particular project has the ability to assign the “Project Member” role for that project to other users, which in turn grants them the ability to see everything in the Project’s Digital Library and Versioned File Spaces. Most importantly, a user may assign roles to other users and then give different kinds of access to their personal data. This allows a user to maintain separate lists of friends, colleagues, and collaborators, each with different abilities. Perhaps friends are allowed to see the contents of the user’s digital library, colleagues are allowed to read the user’s notes within the contents, and collaborators are allowed to add notes of their own in response.

3.2.2 Solo Individuals

Requirements for a user working alone cover some of the most basic tasks. Users must be able to add and edit — whole or in part — knowledge transfer artifacts

in the system. The system should be agnostic with respect to the data stored in the artifact, as well as its format, though common formats should have support for in-line viewing (PDF, JPG, AVI, etc.) and possibly editing (plain text, HTML, RTF, etc).

Once in the system, users should be able to associate notes with the artifacts. The system should automatically associate a timestamp with each comment, and ideally include a digital signature of a sufficient strength to support patent applications for users treating the system as an inventor’s notebook. The user is also able to edit the notes associated with an item, in which case the system will preserve the history of the notes, allowing the user to examine the revision history of each edited note. Notes should be searchable, both for finding an individual note, but also for finding the item the note is associated with.

The items will also have multiple keyword tags associated with them. These tags can be used to help the user quickly find the library items, as well as items that are related. Users are also able to order the tags into a hierarchy, which is then treated as a browsable structure, much like a file system; unlike a file system, however, artifacts can appear in multiple parts of the hierarchy at the same time. At each level of the tag hierarchy, users will be able to specify their preferred order of the tags; this order will be used as the default “natural” order when sorting artifacts at that level.

3.2.3 Mobile Individuals

There are three types of mobile users to consider. The first is a user whose primary work environment is mobile, such as a laptop computer; for the purposes of

this discussion, these users are no different than users with a single immobile work environment. There are also users who have a primary work environment and one or more auxiliary work environments. These may be secondary computers, a laboratory station, or a home office; secondary computers in this case may be a mobile platform such as a smart phone or tablet, instead of a traditional computer or laptop. Finally there are users who do not have a personal auxiliary work environment, but must instead use a borrowed computer or internet cafe to access their data.

Once mobility is added to the user's needs, it becomes clear that the primary system should be on some kind of internet server. In addition, there needs to be a method of syncing the library information to the user's current work computer so that it may be taken offline, perhaps on an airplane or to a field station without internet access. The variety of possible computing devices – laptops, smart phones, internet cafes – used to access the system suggest that full access be possible using a standard web browser; this may be the user's primary interface, but it does not have to be. This variety also suggests that multiple user interface styles should be supported, so that native clients for a range of devices are supported in addition to the web-based interface; we expect native clients will be faster and more convenient, and would include the ability to be taken offline and automatically sync to the main system once they are reconnected to the internet. Users should be able to request full encryption between their current interface and the digital library; for example, when using a web browser, users should be able to log in and interact with the system using the HTTPS protocol, and native clients should incorporate a similar security measure.

3.2.4 Interconnected Individuals

People rarely work entirely on their own anymore. Most work is done in the context of a larger organization, such as a laboratory, business group, or other project structure. While each user wants to maintain their own repository, they also need to be able to share with their collaborators. At the same time, some data in a user's library may be private, so easy to use, but strong, security controls need to be available.

The knowledge transfer system should incorporate a security model based on access lists, user roles, and user relationships. When using role based security, each user is assigned one or more roles, such as project member, project manager, administrator, or guest, and their access to each item is granted or denied based on their role. This simplifies the administration of access rights because, rather than configuring access for each individual, access can be granted (or denied) for all people in a specific role, leaving only exceptions to be handled on a case-by-case basis. By default, all users are considered guests of every other user until they are assigned a different role by one of their collaborators, giving them additional access to that person's library.

For example, Steve is a new user and is not allowed to view Jessica's or Andrew's library. Initially, when Steve tries to look at his co-worker's knowledge artifacts, he is denied access because users in the "Guest" role are not allowed to view either Jessica's or Andrew's data. When Steve and Jessica begin working on a project together, they each add the other person to a "Collaborator" role, and give users in that role permission to view a set of artifacts and their associated notes. After a few weeks, Steve and Andrew have had some time to talk, and they add each other to a "Co-Worker"

role, which gives them both access to view some of each other's artifacts, but not the associated comments.

Users should also be given the ability to allow others, again based on permissions, to add notes to their artifacts, or notes on those notes. This feature expands the basic note association into a full threaded discussion board, but could present problems when a user edits a note someone else has replied to. When a user replies to a note, that note becomes the context of their reply; allowing the original author to edit it would destroy the context for any replies, making the entire thread confusing. To prevent this from happening, in a system that supports this feature, users may only edit their notes by appending to them, as opposed to wholesale rewriting. Additionally, it should be possible for the library owner to flag a note and any threaded replies associated with it, as obsolete or, if an updated note is available, superseded; obsolete or superseded threads are not included in search results unless specifically requested.

3.2.5 Extensible Implementation

The state of the art for web design quickly can quickly outpace almost any web application design. Disruptive technologies are sure to come along every few years; the current explosion of mobile devices and tablets will likely be reshaping our daily use of technology for some time to come. The best way to ensure a smooth transition during such disruptions is to cleanly separate the interface of the implementation from the actual model, enabling multiple styles of interfaces to co-exist; this would include modern web-based interfaces, mobile app and tablet interfaces, as well as interfaces

built into desktop operating systems. Additionally, as new backend technologies are developed (new data storage styles, for example), the model's data services could be migrated with little or no disruption to existing interfaces.

3.2.6 Other Features

The features described above are the cornerstones of the complete knowledge transfer system, but other features could be added to enhance its functionality. User and project blogs would be a good way to enable public outreach and communication. Tags, keywords or short phrases that are descriptive of, or associated with, artifacts within the system, could be a powerful organization and search tool. For reasons noted in Section 3.2.1, not all research can be kept on a system outside the control of the institution sponsoring it, so a cannot assume it is the only installation in the world. Integration between different installations should be enabled via a single-sign-on system that allows users to log in to their home system and still participate in Communities and Projects hosted on other installations without needing a second set of credentials (subject to coordination between the owners of the respective installations). Finally, integration with other forms of social media, such as Facebook or Twitter, would allow users to share some of their work with their friends and the public in general.

3.3 Applications of the Model of Academic Knowledge Transfer

The model of an academic knowledge transfer system presented here covers the gamut of the academic research process. While a complete software system based on the model would take a significant effort to develop, even with a team of people, partial implementations could still be of great benefit. The Mendeley [126] digital library, previously discussed in Section 1.3, is not based on the model presented in this thesis, but it does implement a portion of it.

Mendeley’s major organizational feature is called a Group. Most groups are public and anyone can join; administrative access can be granted to additional members by the group’s creator. Private groups may also be created, but are capped at 50 members for a paid account (only 3 members are allowed with a free account). In relation to the knowledge transfer model, private groups are most like projects (described in Section 3.1.3) and public groups are most like Community Discussions (described in Section 3.1.4). All groups in Mendeley support discussions, which can be associated with a specific document in the group or posts to the group. Groups also support collaborative markup of PDF files and other formats (as supported by Mendeley’s plugin system).

While the groups in Mendeley provide a powerful feature set, the model developed in this thesis suggests several areas for improvement. For example, there is no way for a group administrator to curate the documents added to the group — any member may add to it. A curation process would enable a public group to act as an

electronic journal, with the group’s founder and administrators acting as editorial board and review panel. At the user level, while its possible to create a profile and link to your publications (and see how many people view and download papers you’ve authored), there is no mechanism to post status-like messages, or author documents that are public but not associated specifically with a group. If folders within a user’s library could be made public, documents within the folders could act as a blog published by the user. While Mendeley provides great support for managing documents, it has limited capabilities around user interaction (comment-based discussions in groups and private messages).

In addition to the basic features of a digital library and social network, Mendeley also supports a public API which, as noted in Section 3.2.5, is useful for both future proofing the implementation and enabling support across multiple platforms. This has allowed Mendeley to offer a native client application for both desktop and mobile clients, in addition to its web client. The native apps can sync with Mendeley’s servers when they have internet connectivity to get the latest comments in a group or on a library item. Some features, such as adding annotations within a PDF file (as opposed to comments associated with the artifact), are only supported in the native clients.

Chapter 4

Whisper: A Prototype Personal Digital Library for an Academic Knowledge Transfer System

Along with the Knowledge Transfer model presented in Chapter 2 and the general system design described in Chapter 3, a prototype implementation of a major aspect of the environment can provide insight into design details and user interest. The prototype described here focused on the digital library feature of the overall system. This feature was selected because it would be useful to each entity described in Section 2.1: Individuals, Project Groups, and Communities. Most importantly, individual users could gain some benefit to using the prototype even if they had no social network within it. While no more important to the overall knowledge transfer system than versioned file systems for projects or peer-reviewed electronic journals for communities, the digital library feature can act as a gateway into the overall system, were it to be completed.

4.1 Introduction

Academics have had access to digital copies of published work for decades now, but the web has dramatically increased the creation of digital libraries by both publishers and professional societies that sponsor research. And once built, academics have flocked to these new online resources as a convenient source of research material. However, the dark side of this trend is the large numbers of digital articles stored on each researcher's computer (or computers). Without a personal digital library capable of managing the flood of papers and other research material, each individual has had to manually organize the files they have collected, along with any related notes; this usually leads to hours of work either maintaining the organization or searching a disorganized file system. Often, institutional digital libraries provide "personalization" features, such as alerts or social networking tools, making it easier to find new material or revisit articles uncovered by previous searches in the library, but these features fall far short of providing a truly personal digital library.

The ACM has a well known digital library, used by researchers in Computer Science and related fields from around the globe. There has been a significant amount of both research and development done for this style of institutional digital library, but features for individual users are still light. For example, the ACM Digital Library [2] has a feature for individuals called "My Binders". Users are able to manually add entries or create a saved search over the content of the entire library; however, users can only add content available on the ACM Digital Library. Users can also share their binder with other individuals, or groups made up of the various ACM SIG memberships; there is

no way for a user to create a group of their colleagues and share their binder with that group, they must instead add their colleagues individually to each binder they wish to share.

In this chapter, we present an alternative from the overall Whisper system, focused on the needs of the individual users. We built our requirements by first examining the needs of a user working by themselves at a single computer, then expanded their environment to include secondary work areas and working on-the-go, and finally incorporated the need to work with others. We also describe the architecture we used to create a prototype personal digital library based on these requirements, and our experiences building the prototype and testing it with users over the course of a school term.

The chapter is organized as follows. We first examine previous work around personal digital libraries and their architectures in Section 4.2. Section 4.3 presents a set of requirements for a personal digital library. In section 4.4 we describe a software architecture capable of meeting those requirements. Our prototype implementation of this architecture is described in section 4.5, followed by our experience developing and using this prototype in the classroom environment in section 4.6.

4.2 Related Work

Gonçalves et al. developed a formal model for digital libraries [75] which was later extended by Ma et al. to support a personal digital library with personal information management functionality [121]; UpLib [94] and MyLifeBits [70] are personal

digital library systems with personal information management features that have been in development for several years. These projects include features for capturing personal information from the user’s daily life, including receipts, digital photos, and other types of digital media, with the goal of keeping most, if not all, data from the user’s life in a searchable digital repository. The scope of these projects is larger than our vision of a personal digital library, which contains data and research materials but not the bill from this week’s dry cleaning. We feel it is better to separate personal information and research materials into different repositories because the context for their use is so different; when searching for research material relating to Spain, users do not want their search results cluttered by photos of their recent trip to Spain or recipes for Spanish food. In contrast, commercial software, such as Delicious Library [48], generally does not provide a large enough scope to be an effective organizational tool for research material. Users are limited to organizing their media library and bookshelf, and there is no support for the inclusion of “partial works” such as articles, audio/visual clips, and individual images.

Alvarez-Cavazos et al. developed an architecture similar to the one described here for digital libraries [5] and have created a system called PDLib based on that architecture, incorporating many of the features we describe here. The internal organization of documents within PDLib is modeled after directory hierarchies in a file system. Unfortunately, their description does not include enough detail for us to determine whether users are able to keep the same item in multiple locations, a feature we believe is critical because so many references are used in multiple projects; additionally, while multiple

users have accounts within a PDLib installation, there are no collaborative features described other than simple document sharing. A large portion of the PDLib server architecture is focused on ensuring high quality access for mobile devices, especially low powered and resource scarce systems; we feel these needs are more appropriately handled at the client level because they are specific to a particular class of device.

Chen et al. also developed a personal digital library architecture [35] which splits the data management layer into separate systems for file, metadata, and policy storage, and also provides for content authoring and editing within the digital library system itself. Our view of a library does not include unfinished works, other than notes on the items, being authored by the user – version control systems or content management systems are a better fit for that type of data.

Hicks and Tochtermann [84] developed an architecture that stores the user’s metadata locally, but allows the digital representations of the library items to reside in their original location (such as the ACM Digital Library, or a NASA data archive); when a local client requests a digital representation, it is first retrieved from its original location and then customizations (filters, cropping, image overlays, etc.) defined by the user are applied. Our concern with this approach is its performance when the user is offline, or worse, if the original items the user has captured in their system are moved or deleted by the organization that maintains the external site; except in rare circumstances, such as a library of feature length movies, we believe users are better served if they are able to maintain a private copy of the items in their library.

The system we describe here focuses on providing a personal library for research

materials, rather than a media library [48] or a personal information management library [121, 94, 70]. We also provide for storage of the library item data within the system, rather than forcing it to be hosted externally [84]. Finally, we provide a strong feature set that integrates personal needs with group collaboration, while also supporting multiple client applications.

4.3 Digital Library Requirements

The requirements for the Digital Library prototype build on the general requirements for users presented in Section 3.2. Some of these requirements have been met for years with both free and commercial tools, such as BibTeX [16], End Note [52], and UpLib [94]. Library items consist of bibliographic metadata about the item, including title, authors, publication information, and permanent location where the original item may be accessed if the user's copy is lost (see Table 4.1 and Table 4.2 for a complete list of supported metadata); the permanent location may be a physical location, a web location, or some other location data as determined by the user. The user is not expected to enter all fields of an item's metadata, but the title is required for each item, and a family name or organization name is required for each author. If the user has a digital representation of the item they should be able to associate that representation with the library item inside the system; when possible, these representations should be indexed to support content-based search within the library. Digital representations include:

- A PDF or Microsoft Word file for a paper, interview transcript, or book.

Table 4.1: Library Item Metadata

Title	Sub-Title
Authors	Tags
Date Created or Published	Periodical Title
Volume	Number
Issue	ISBN or ISSN
Start Page	End Page
Publisher	Publisher Location
Abstract or Description	Source URL

- A JPG or TIFF for an image, painting, or sculpture.
- An MP3 or MPG file for an audio or visual performance.

In a system with more than one user, there is no reason to force each user to enter the bibliographic metadata of a library item if another user has already added it; in general, most of this data is public and freely accessible, even if the item it describes

Table 4.2: Author Metadata

Salutation	Family Name
First Name	Other (Middle) Names
Suffix	Organization Name
Title	URL
Email Address	Physical Address

is private or protected by copyright. At the same time, users may have occasion to enter a new item before it has been published, in which case sharing any of the metadata might not be appropriate. Because of this, when entering a new library item into the system, users must have the ability to flag the metadata as private, which will prevent it from being shared with other users; users should be able to remove the flag, but once removed, it cannot be restored.

4.4 System Architecture

Our system architecture (see Figure 4.1) is based on the classic Mode-View-Controller pattern [30]. The Model is the Resource Tier, which manages the raw data using a database or other type of content repository. The View is the Client Tier, with support for both web and native clients. The Controller mediates the flow of data between the Model and the View, with high-level services handled by the Services Tier and lower-level services handled by the Data Management Tier.

4.4.1 Resource Tier

The Resource Tier contains all data stored by the users of the library system. Most data is best stored in a standard relational database, though some types of data may be better stored in a different area, perhaps even the file system; for example, it might be inefficient to store the digital representations of the digital library items in a relational database. User profile data is probably best stored in a database; this profile data includes names, passwords, mailing address, research interests. Social

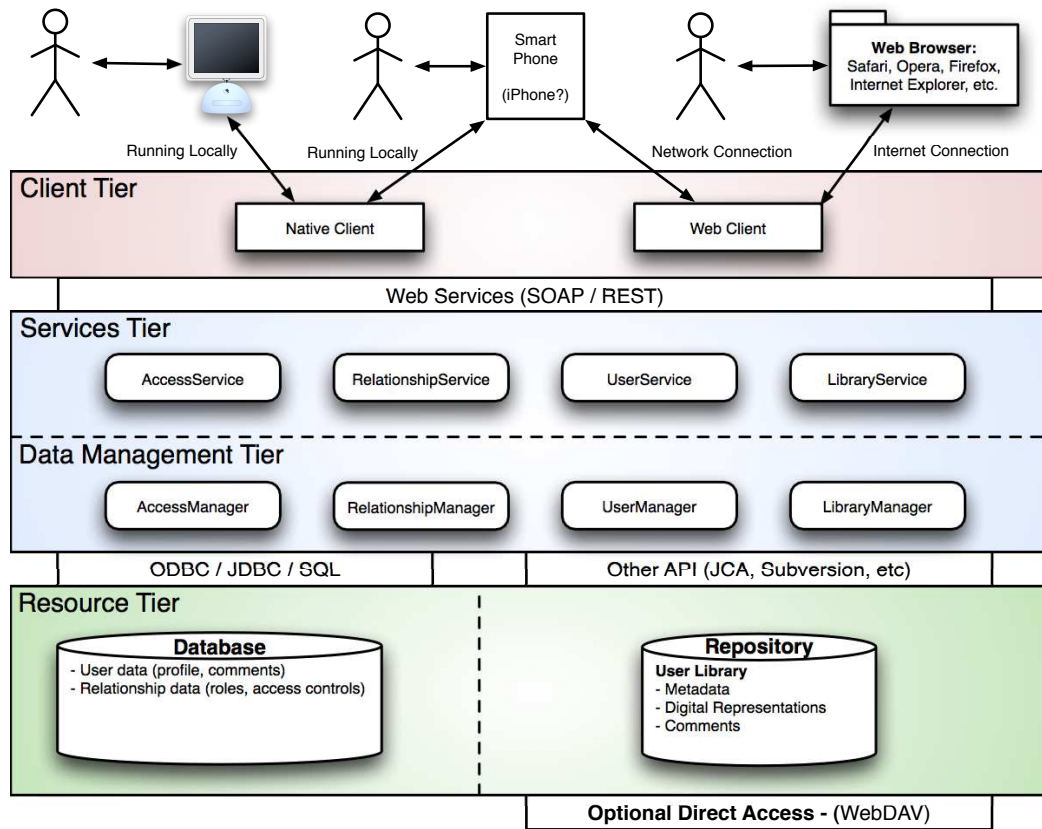


Figure 4.1: Conceptual Architecture

networking data, such as the group and role membership lists, is also appropriate for database storage. The library item metadata could be stored in the database as well, but implementations could reasonably consider other storage systems such as a Subversion [150] repository or dedicated Content Repository, such as one conforming to the Content Repository for Java API [98]. These other data storage methods provide some additional features, such as versioning and could be configured to allow direct access over the web, or be treated as a read-only network attached file system; while none of

these are requirements, they are certainly useful features.

4.4.2 Data Management and Services Tiers

The Data Management Tier contains the software components that interface directly with the Resource Tier via ODBC, SQL, HTTP, or another appropriate protocol. Each component represents a basic abstraction within the library system. The UserManager provides access to each user's profile data, and the RelationshipManager maintains the role membership. The AccessManager handles the role-based security and works with the UserManager and RelationshipManager to determine who has access to specific resources within the system. The LibraryManager maintains the content of the library items and their digital representations. After retrieving the requested data from the Resource Tier, the Data Management Tier filters the query results, removing data the requesting user does not have permission to view; during write operations, the user's permissions are checked before changes are made to the stored data.

The Services Tier gets the permissible data from the Data Management Tier and packages it for the Client Tier. In some cases, this may mean generating a BibTeX file, or transmitting raw data to the client. The Services Tier is also where any data transactions needed to process a request should be created and maintained, since any one request may span multiple Managers. Implementors may also use the Services Tier to provide WebDAV [74], OAI Harvesting [108], or other methods of access to the library system, if those methods are not supported directly by the Resource Tier.

4.4.3 Client Tier

The Client Tier is somewhat different in that it represents any number of different systems. The Resource, Data Management, and Services Tier are all contained within a single entity (even if that entity resides across multiple machines for scalability). The Client Tier interacts with the Services Tier using a form of web services, such as SOAP [78] or REST [65]. This architecture will support multiple styles of user interfaces, including web-based clients built with PHP, Ruby, Python, or any other language with support for the chosen style of web services. Native clients are generally hardware or operating system dependent and may be written for desktop or mobile computing environments; these clients can be expected to cache a portion of the user's data on the local system and sync with the main library system.

4.5 Implementation

Our prototype, seen in Figure 4.2, was developed by a single graduate student over the course of a year. The user interface was designed before development of the code began. Since programming resources were scarce, some of the technology decisions were driven by a need to reduce the complexity of the code. The user interface was written with PHP, and the rest of the system was developed using Java. The various layers of the server-side Java application were tied together using the Spring Framework [148].



Figure 4.2: Screenshot from the working prototype

4.5.1 Data Model Implementation

Following our architecture, the data model was implemented in two parts. The core of the system was stored in a database, and the library items were stored in a Content Repository, which relied on a separate database on the same server for storage. Each Manger component was split into two classes; a Manager class provided the logical data management, while a Data Access Object (DAO) class provided the low-level functionality dictated by our choice of data management technologies.

Rather than using raw SQL to marshal data in and out of the database, our prototype made use of the Java Data Objects (JDO) specification, which allowed us to design our object model normally, and create files that mapped the classes and their relationships to database tables. The result was a data model that performed well

enough for use in a small-scale prototype system, without the need to create custom SQL statements beyond simple filters; were this prototype to be widely deployed with thousands of users, we expect some amount of database and SQL tuning would be required for adequate performance across the entire system.

We choose the Apache Jackrabbit project [92] for the content repository that stored library items and their digital representations; Jackrabbit is a complete implementation of the Content Repository for Java API (JCA) [98] and very close to its 1.0 release when we began working with it. The content repository consists of nodes with defined properties. Some properties are simply strings or dates, while others are references to other nodes; where appropriate, multiple instances of a property are allowed. To store and retrieve data from the repository, the `LibraryManager` makes calls using the JCA API, marshaling the properties from Java objects to the repository properties and back.

4.5.2 Controller Implementation

Our choice of the Spring Framework made the work of knitting together the components within the layers a relatively simple task. By taking advantage of the aspect oriented features provided by Spring, we were able to configure security, transactions, and web services in XML files, rather than writing a large quantity of boilerplate code. When a request comes in, before any method in the Services Tier is invoked, a Spring-based interceptor ensure that an appropriate database transaction is in place to handle the request. This transaction is propagated down the call chain and either commits or

rolls back when the Service Tier's method finishes executing. Between the Services Tier and the Data Management Tier, additional Spring-based interceptors ensured that the user making a request had sufficient privileges to call an operation, and if the operation was allowed to proceed, a similar interceptor could be used to filter the results so that the user did not see data without permission.

For example, an anonymous user (someone who has not yet logged in) would be allowed to request a list of other users and the list would be returned without filters; that same user would be denied access if they attempted to edit the list of organizations one of the system's users belongs to. Anonymous users are also allowed to request user profile data for all users in the system, but before that list reaches the Services Tier, it is filtered to remove the profile data of users who have restricted access to their profile data; the same process is performed when known users request profile data.

Our prototype implements SOAP-based Web Services, using the XFire [173] implementation library because it worked very well with the Spring Framework. We also found XFire to be very easy to configure, and it allowed us to generate the Web Service Description Language (WSDL) files based on Java Interfaces; this led to a more natural development process compared to generating Java classes based on hand-crafted WSDL files.

Just as the Manager components were segmented into two classes, the Services components were as well. The bulk of the logic is implemented in a Services class, while the web service implementation specific code is in an RPC class. In our implementation, XFire builds the WSDL files based on the RPC class interfaces. Each concrete RPC class

contains a reference to its associated Service class, and uses Java Annotations to specify web service options, such as parameter and result value names; when retrieving digital representations of library items, the RPC class converts the internal representation of the data to an array of bytes for transmission to the client. The Services class itself retrieves the user from the request and ensures all the request parameters are present before calling the underlying Manager.

4.5.3 View Implementation

The user interface for our prototype is written in PHP, using the SOAP implementation from the PEAR library which we found to be compatible with the XFire-based SOAP interface, while the built-in PHP SOAP implementation was not. The design of the interface was completed before dynamic pages were popular, and even expected, by users; our decision not to update the design with more dynamic features was the source of much of the user feedback we received, as we will discuss later. Our data model in the UI mirrored the object types and services provided by the server-side implementation. Most of the work done by the UI simply involves marshaling data in and out of SOAP messages. We also provide the user with a JavaScript-based widget for editing the keyword hierarchy; this was the only case in which we implemented dynamic functionality within a web page because it was obvious that the user experience would be significantly hampered without it. The UI did not locally cache data from the back end of the library system; in a full-scale deployment, some amount of data caching would probably be required.

4.6 Lessons Learned

4.6.1 Implementation Choices

Our decision to use JDO to store our model objects in the database was driven by its status as an open standard, and the availability of both commercial and open source implementations. The other major Object Relational Mapping (ORM) solution at the time was Hibernate. As Hibernate was more popular and the tools more developed, even though it was not a standard, the development process would have been a bit smoother had we chosen it over JDO. Since we made our choice, Hibernate has morphed into an implementation of the standard Enterprise Java Beans 3 Persistence API, which would have met our desire to use standards-based ORM solution.

The larger drain on developer resources in our prototype was the use of the second technology, the Apache Jackrabbit implementation of the Java Content Repository API, to store the library item data. The amount of code needed to marshal data between the JCR repository and the internal model of the library item data required almost as much code as all other data access code in the system. As we added attributes to the library items, we were required to manually update the data access code in several places, rather than simply adding the new field to a configuration file. In addition, JCR repository itself was, compared to JDO, a heavyweight layer to put between our internal data representation and the database. While we did not run any performance tests, individual integration tests around library items took significantly longer to run than similarly scoped tests against JDO based objects.

4.6.2 User Feedback

Our prototype was used in a classroom setting for a seminar class during one school term. Students were required to read several papers each week and make notes on the papers within the library system; the notes were then reviewed by the professor. Towards the end of the quarter, we asked the students to share their experience using the prototype with a third party, versed in Human Computer Interaction (neither the course professor nor the prototype's developer were present to encourage the students to speak freely). They first took part in a freeform discussion of their experiences using the system, followed by a form of participatory design in which they were asked to sketch their ideal version of the user interface.

While the students felt the prototype had the potential to be an excellent tool, much of their feedback reflected its unfinished state; many of the major features presented here, such as searching and sharing, were either incomplete or unavailable to the users. In addition, for features that were available, common actions such as altering the sort order of a list of library items were not possible. Another common complaint was that navigating the site required too many clicks and page loads; we believe the source of this problem to be the lack of dynamic content loading, as is now common for websites and expected by users. The users also felt that the pages tried to pack too much information into them, and that the quantity of information was overwhelming; this is another area in which the interface would be improved through the use of dynamic content. Some users suggested integration with external sites, such as Google Scholar, would be helpful. The interface designs suggested by the users also reflected these

criticisms, with many of the drawings using mouse-over driven popup elements, such as item details, commands for the current element, or search areas and lists of related content.

4.7 Interface Alterations

A significant critique of the Digital Library system implemented as a test bed for Whisper (other than missing features, which is an expected criticism for prototypes) was that it was hard to use. In part, this was due to the UI including placeholders for future sections of the Whisper system. But a larger part was the fact that pages were static, meaning that once they were loaded, almost any click would result in a full page reload, rather than a dynamic action (via an AJAX call).



Figure 4.3: A sample page from the digital library test bed.

As seen in Figure 4.3 (the same image as Figure 4.2), the static nature of the pages required a large amount of the page dedicated to navigation and issuing commands. The top of the page contained the primary site navigation, allowing the user to view their Home (where their personal data would be), other People (friends and collaborators), their Projects, and their Communities; there were also tabs for Search, Preferences, and Help. The column on the left was dedicated to navigating within the tab that was currently selected, and a column on the right had the available commands for the current view. This left about 50% of the width of the page for the actual content the user would be working with, which leads to a very cramped area. The only dynamic aspect of this page is the small arrow icon that hides and shows item-level commands for the entries in the library. Since most of the site's pages would have considerable amounts of text, this design makes it hard to space that text out and separate it in a way that is easy for the user to skim and comprehend.

At the time the site layout was designed, dynamic web pages with heavy use of AJAX were still new. But during the development of the prototype, the dynamic, Javascript-driven web design became very popular and many high traffic websites (such as Google Maps, which helped pioneer the techniques) were used often by the students who participated in the study. Even sites with less dynamic interfaces than Google Maps had adopted some AJAX techniques, like in-place editing (replacing static text with input fields, and then switching back to static text after the fields were edited). The Whisper prototype had none of that.

One of the activities the users were asked to do, as part of providing feedback,

was to suggest alternative designs for the application; many of the designs included dynamic sections of the pages. After examining the feedback, as well as the overall information content of the prototype, a new page design was devised that both allowed for more content to be at the user's fingertips, giving the page's main content more space and hiding most commands in large popups. The resulting design (seen in Figures 4.4 through 4.6) includes a header at the top of the page, followed by four tabs for the major divisions of the site, and leaves the rest of the page for the current content.

In the header, there is room for the site's logo, a general search field, as well as links to site help, and user settings. The four tabs – My Work, Contacts, Projects, and Communities – respond to clicks by opening a large navigation panel related to that area of the site. Each panel is several hundred pixels wide, allowing three columns of short links or commands; this provides space for both static links to important areas within that area (perhaps a list of the user's current projects), as well as context-sensitive links for the current page's content (such as related papers, recent comments, or a project's collaborators). The page content can be divided between one and three content areas, optimized for the type of content. For example, search results (diagramed in Figure 4.4), can use a pane in one half of the content area for the list of results, a two-thirds height pane for the search and sort criteria (which may be editable), and a smaller one-third height pane for details of each item as the user hovers the mouse over it in the results list. Three content panes would also be useful for viewing the details of items within the digital library, with the bibliographic data displayed in the large left-hand pane, comments in the two-thirds height pane, and related papers in the one-third height

pane, as diagramed in Figure 4.5. On the other hand, reading a paper in one of the community journals would be best done in a single, full-width column (not diagramed). Figure 4.6 diagrams the open “My Work” tab, with possible data, commands, and links that could be made available to the user.

This new design was prototyped to test its technical feasibility. While the prototype is still a work in progress, the following screen shots give an idea of how the above diagrams might translate (roughly) into an actual page design. Figure 4.7 shows a basic page with an open “My Work” tab; implemented links allow the user to view all items in the library and add a new item to it, and provides a short list of recently added items and comments. Figure 4.8 shows the basic structure of the three panel content design. The main pane on the left, as well as the pane on the top right column will both expand as the browser window grows, while the pane on the bottom right panel has a fixed height. Static text tends to be much more compact than the HTML interfaces for editing it; by breaking the data being edited into smaller groups, an accordion widget (as seen in Figure 4.9) can be used to show only one group at a time, while making it easy to move from one part to another.

4.8 Conclusion

We believe there is a strong need for digital library systems designed primarily for individual use where users are able to collect the breadth of their research work into a single system and share that work with others. Oleksik et al. noted, “further progress is dependent on supporting not only data access but the entire process of scientific inquiry.

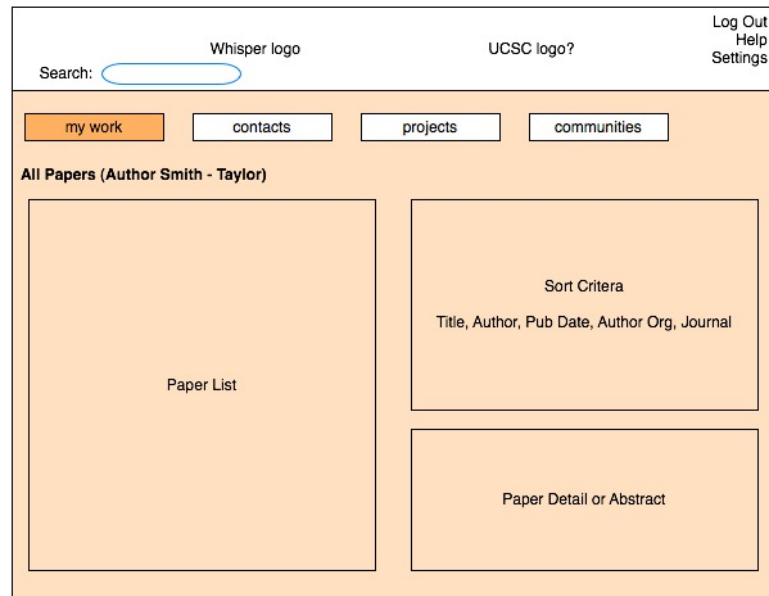


Figure 4.4: The redesigned digital library screen.

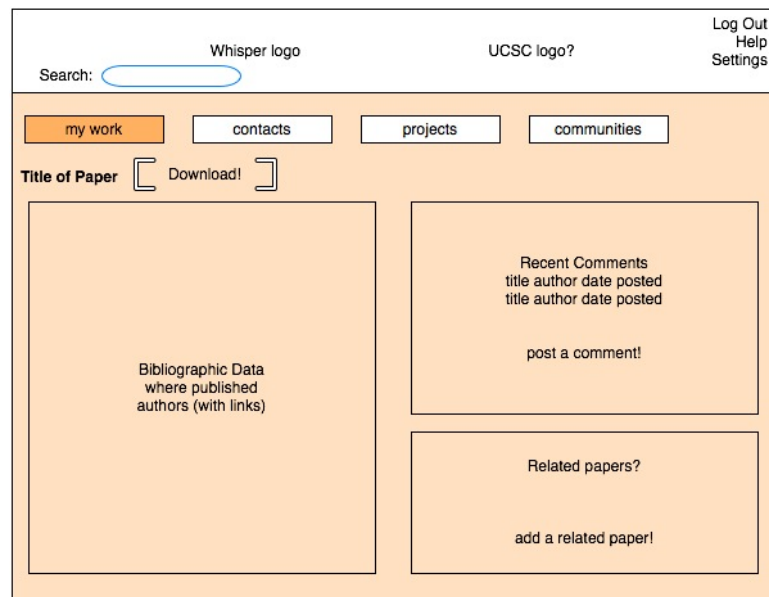


Figure 4.5: The redesigned item details page.

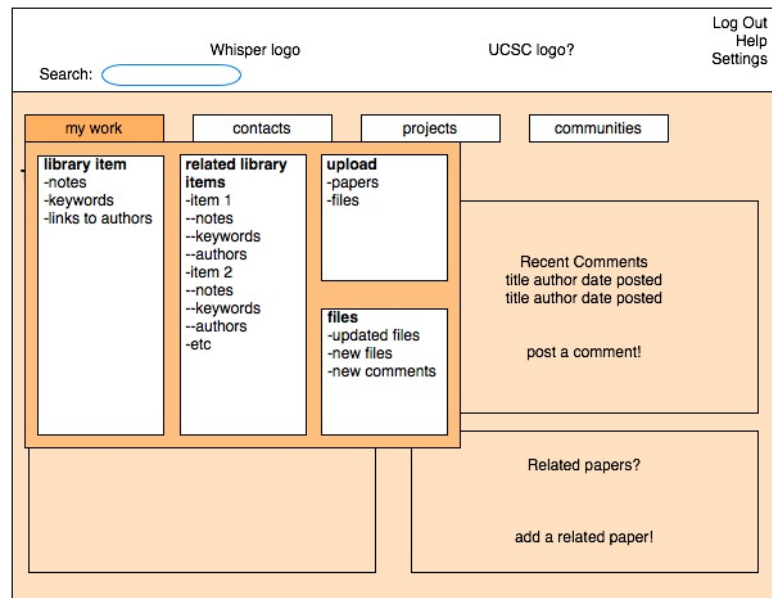


Figure 4.6: The redesigned item details page showing the open “My Work” panel.

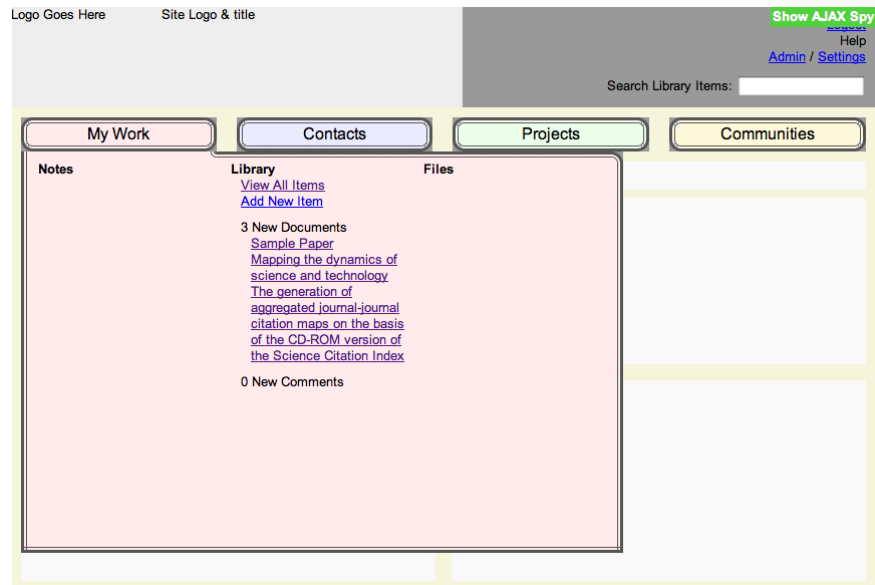


Figure 4.7: A prototype item details page.

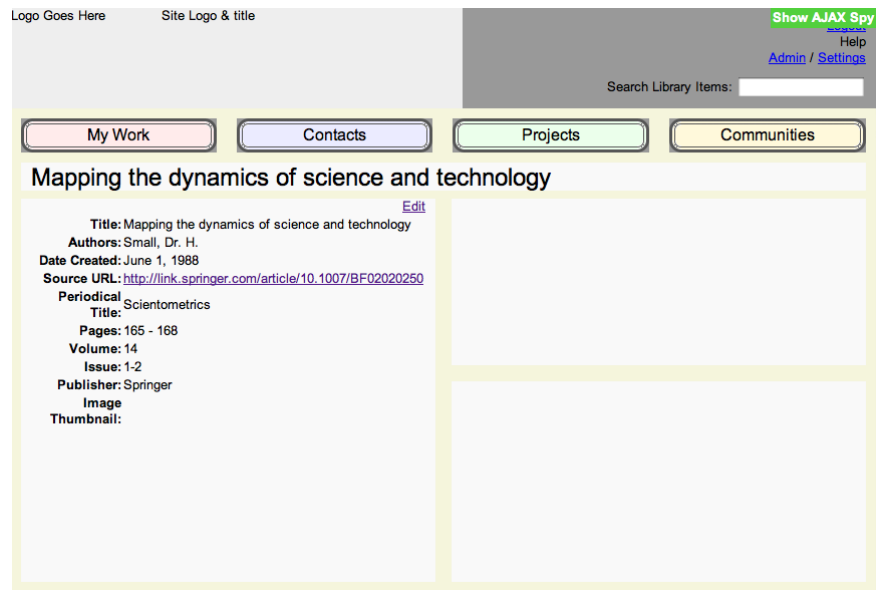


Figure 4.8: A prototype details page showing the open “My Work” panel.

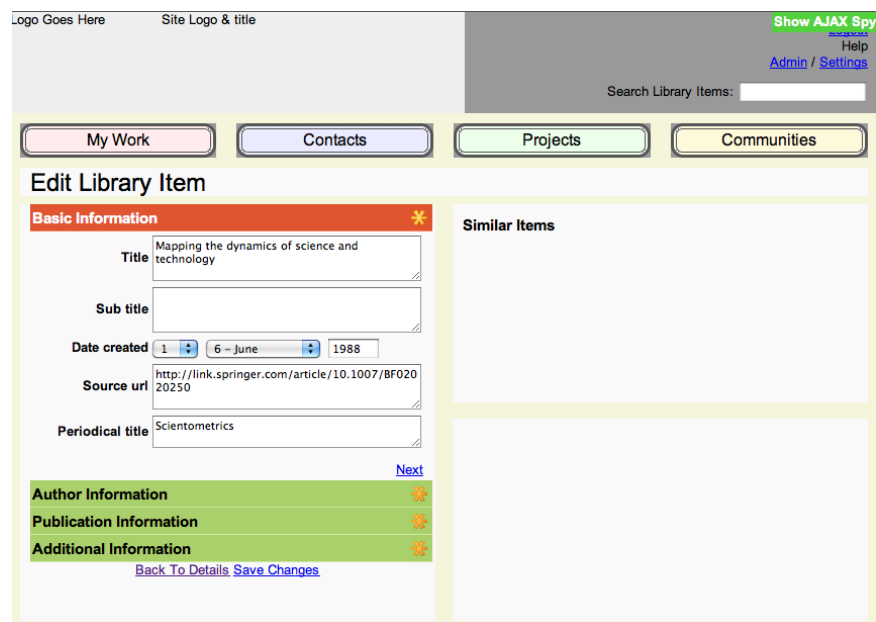


Figure 4.9: A prototype edit page.

That includes a range of individual and team practices, from running experiments to preparing scientific publications.” [131] The features and architecture here are, perhaps, a minimalist place to start; once there are systems of this nature in common use, listening to users will certainly give rise to more features than we have considered here.

We have presented a comprehensive feature set for personal digital libraries at three levels of usage: individuals at a primary workstation, individuals with multiple (potentially mobile) interfaces, and individuals within the context of their larger organization. These features include a novel tag-based management system that allows users to efficiently organize and browse their digital library. The feature set also includes strong security between the user interface and the back end system, and role-based access controls when multiple users are sharing the same system; these security features provide more flexibility and control than we have seen described or used in other personal digital library systems.

We have also described an architecture capable of supporting many styles of user interfaces, including web interfaces, and native interfaces for both mobile devices and desktop computers. A prototype we developed based on this architecture showed the potential to be valuable tool for the students who used it, even though its feature set was very limited at the time. The feedback we received provided very clear directions for user interface modifications in our system, with users having come to expect a more dynamic experience than our prototype provided.

Chapter 5

A Cautionary Tale of Turning Work Into Play

After the prototype application to explore the Digital Library feature of the model and software system was developed, it seemed prudent to explore the development of the software system within an existing, and very popular, social network. The expectation was that users of the system could be enticed to share links to academic knowledge artifacts, such as published articles and news reports, using various game-based feedback mechanisms; a process commonly known as Gamification.

5.1 Introduction

There is a general bias in the games community that suggests any work task can be made fun, thereby encouraging people to participate more often or more fully. Webster and Martocchio [167] have shown that simply calling a task play instead of

work has a significant impact on some people's willingness to participate, especially younger people. Some people, such as Wagner in his article "Turning Work Into Play Is No Game" [161] espouse using virtual worlds like Second Life to make repetitive office jobs bearable and to allow remote employees to feel more connected to their coworkers through virtual hallway encounters. Others, like Reeves and Read [138] suggest experiences in online gaming can develop useful qualities like leadership and negotiation just as effectively as real world situations.

There have been many studies [38, 50, 110] of social networking on college campuses and a few studies in the workplace. Ellison et al. found that weak ties between students were strengthened through the use of Facebook [50], while Lampe et al. found that students don't use Facebook to make new contacts as much as to learn about people they met offline [110]. In a study of lecturers, Colete et al. found that very few were making use of Facebook as a learning tool; we found this in our research as well. DiMicco et al. discussed the success of their Beehive [49] project at IBM; employees developed stronger connections with people they rarely got to see in person, and across all levels of the organizations. But most research on turning how to turn work into play has been focused on training [43, 161].

However, there have been no studies showing how much fun is needed to see more user engagement, and perhaps if too many elements from games might instead discourage users by drowning out the task at hand. The original goal of this research project was to fill some of that void by taking two related tasks common in academic research and building several applications to support them, each one with a different

level of gameplay elements. We chose to use two common activities from academic research, sharing background research with other interested friends and collaborators, and organizing personal notes related to that research. There are many existing methods of sharing such information, email being the most obvious, but emails can easily get lost in the flood of messages received each day. Additionally, keeping track of notes for so many resources can be a challenging problem for even the most organized person. We hoped to address these problems by creating an environment in which users could easily share links to any online resource with their friend and collaborators, while also maintaining the notes they made on that resource.

We built three information sharing applications that ran within Facebook: Tidbitz, Nuggetz, and Read All About It. All three had the same basic features: add a resource link, edit notes on that link, find links and notes, and share links with others. Tidbitz is a bare bones application that simply implements the application features, while Read All About It has a print newspaper theme to its look and additional features to encourage users to compete with each other for “readership” and “quality”; Nuggetz falls between the two by giving users feedback points, but without competition based on those points. We expected to find differences not only in how fun affected people’s opinion of each application, but also differences in how people of different genders or cultures reacted to the game-like elements. Instead, we found ourselves struggling to get enough users to do any kind of useful analysis and after two months went about the hard process of identifying where we’d gone wrong.

Pinch and Bijker’s “Social Construction of Technology” (SCOT) [133] offers a

good place to start. While society and technology can be studied individually, SCOT argues that a technology cannot be fully understood without also examining the social context in which it exists. Since this chapter describes only one attempted solution, we focus on the first stage of a SCOT analysis, which identifies the social groups, explores of the problems they face, and examines how the technology at hand attempts to address those problems. This chapter discusses the following questions concerning the lack of adoption of the information sharing applications.

1. How do the relevant social groups (graduate students, professors, and so on) currently share information and manage their notes?
2. Were the applications functionally complete for the tasks they were designed to assist with? Do they solve the problems faced by the people in our social groups?
3. What role did the Facebook environment play in the lack of adoption?

The chapter is organized as follows. Section 5.2 describes the features of the three applications we developed. In section 5.3 we describe how data was gathered from the users, how users were recruited, and the demographics of the user base. A discussion of the survey results and user feedback follows in section 5.4, followed by the lessons we have learned from this experience in section 5.5.

5.2 Study Applications

Three applications were developed on the Facebook Platform [62] in this research project, Tidbitz, Nuggetz, and Read All About It. All three applications are

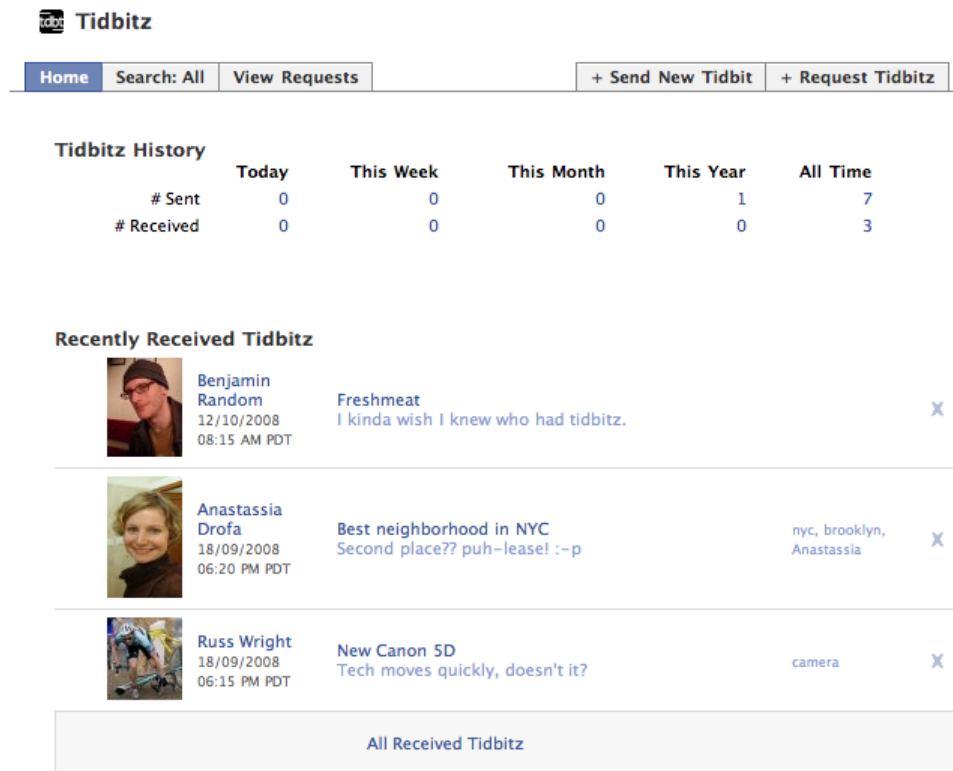


Figure 5.1: Tidbitz Home

built around the concept of sharing information with others and provide the same basic functionality. The primary use case in the research applications is sharing a link to a web-based resource such as a news article, white paper, research paper within a digital library, video, and so on, but there was nothing to prevent users from sharing comics, recipes, or anything else with a URL.

Filter Tidbitz

Search:

After:

Before:

With Tag(s):

Note: Tags are comma separated.

Sender:

Contains Subject Text:

Has No Comment: ☐

Contains Comment Text:

Contains Link Text:

Figure 5.2: Tidbitz Search Filter Dialog

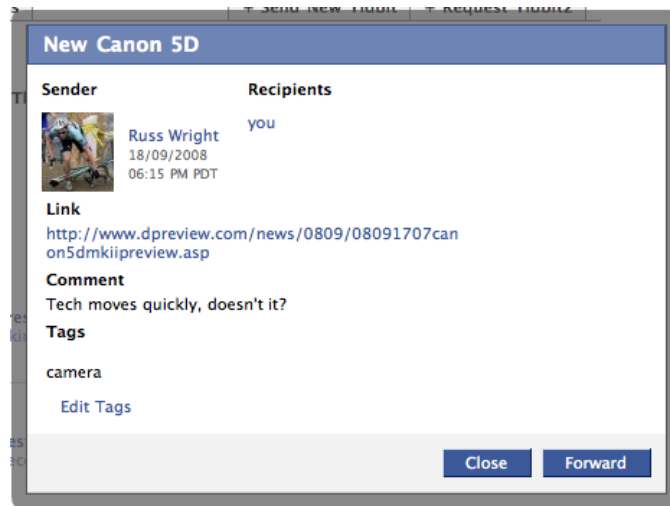



Figure 5.3: Tidbitz Read Message Dialog

5.2.1 Tidbitz

The most basic application is called Tidbitz. Tidbitz has no feedback mechanisms (fun elements) and was designed to provide a baseline for our original research questions; user engagement is driven entirely by its functionality. There are five pages the user can access: Home, Search, New Tidbit, Request Tidbitz, and View Requests.

The Home screen, seen in Figure 5.1, provides an overview of the user's activity within the application, including counts of all messages sent and received during various time ranges (today, this week, etc). Below that is a list of recently received messages; this same style list is used in the search results. Users can search their messages using a variety of filters, as seen in Figure 5.2; ten results are displayed per page.

Each message in the list shows the sender's details with appropriate links to the sender's Facebook profile, and the date and time the message was sent. In the center

 Tidbitz

Home

Search: All

View Requests

Compose New Tidbit

+ Request Tidbitz

Recipients:

Subject:

New Canon 5D

Link:

<http://www.dpreview.com/news/0809/08091707canon5dmkiipre>

Comment:

Tech moves quickly, doesn't it?

Send

 or Cancel

Indicates required fields

Figure 5.4: Tidbitz New Tidbit Page

of each row is the message subject in dark lettering, and some or all of any comment the sender included, excerpted if necessary. To the right of the subject and comment is the user's current list of tags for that message, if any. The subject, comment, and tags may be clicked to bring up a dialog containing the entire message, as seen in Figure 5.3. The user is able to edit the tags for the message by clicking on the "Edit Tags" link; this displays an editor in the dialog, allowing the user to add or remove tags as a comma separated string. The user is also able to add and edit their private notes related to this message by clicking the "Edit Comment" link.

When the user composes a new message, or forwards one they received, they do so on the "Compose New Tidbit" page, seen in Figure 5.4. The user must enter one or more recipients, a subject line, and the link they want to share. Additional comments are optional; these comments are shared with the recipients and are not related to the user's "private notes". The "Request Tidbit" page (not pictured) has a very similar form but has fields only for a subject and comments that describe what information the user is interested in, both of which are required. To ensure users don't see many stale requests, which might drown out the new ones, they may have no more than five requests active at a time. Once they have reached that limit, they must remove a request before creating a new one.

Users can see the active requests that they have created, as well as those created by their friends on the "View Requests" page (not shown). Tabs at the top of the page allow the user to quickly switch between requests from their friends and requests they have submitted. Requests from friends are displayed much like messages, as are requests

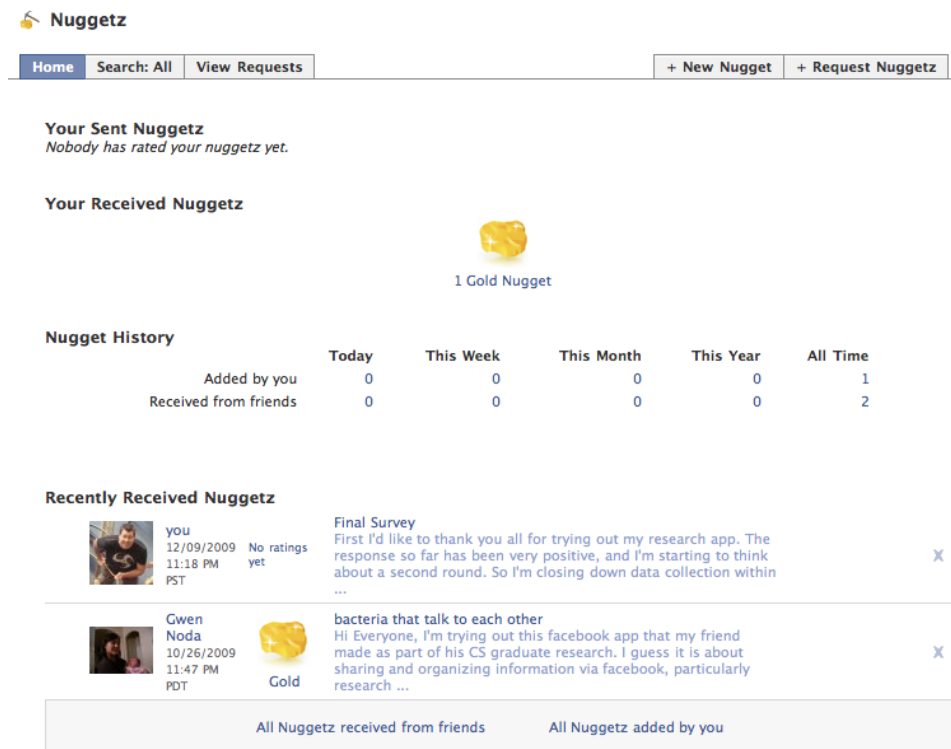


Figure 5.5: Nuggetz Home

from the current user.

5.2.2 Nuggetz

Nuggetz, the second application retains all the functionality of Tidbitz while adding a ratings system for message recipients. The five ratings available, from worst to best are: Toxic Waste, Coal, Bronze, Silver, and Gold. The user interface has only minor changes to display and edit the ratings. On the home screen, seen in Figure 5.5, the user is shown a summary of the ratings given by the recipients to the messages they have sent. This is followed by a summary of the ratings they have given to the messages



Figure 5.6: Nuggetz Ratings

they have received. The message dialog box is identical to Tidbitz's, except for the addition of ratings (see Figure 5.6). Senders will see a summary of all the ratings their message has received, while recipients can select a rating.

5.2.3 Read All About It

The final application provides a more immersive experience than the other applications, making the game aspects more obvious to the users. The users play the role of a journalist, hence the application is called Read All About It. The user interface elements evoke the idea of newspaper articles ripped from the paper to be kept in a scrapbook, as seen in Figure 5.7. The user can increase their journalistic level in two dimensions based on the readership and ratings of the messages they send; the user's friends who have added the application are visible at the right of the screen along with their readership and rating levels. Read All About It also adds a community feature



Figure 5.7: Read All About It Home



Figure 5.8: Read All About It Message Dialog

called the “Global Feed” that allows highly rated messages to be shared with all users in the application.

The readership of a message is determined by the number of recipients, including new recipients when the message is forwarded; readership from forwarded messages only applies to the original message. As the user gains readership points, their journalist persona moves to larger newspapers. All players start at the “Neighborhood Newsletter” level and progress to “Facebook Journal”, the highest level attainable. Each level requires a certain number of messages to attain a minimum readership count and depth, in addition to having a minimum number of rating points. Depth is calculated based on the number of forwards by recipients to other people who are not friends with the original author; this is done to prevent groups of friends from easily gaming the system.

Ratings in Read All About It, seen in the Message Dialog in Figure 5.8, are taken from newspaper sections and count as five to one points, respectively. Rating points are determined by examining the average rating for each message. When there are multiple recipients, the average is based only on the number of ratings that have been given, so if some recipients never rate a message it does not count against the sender. Rating levels require a minimum rating across all messages sent, based on the the user’s current readership depth, and must be from a minimum number of recipients.

The Global Feed in Read All About It allows an original sender of a highly rated message (an average rating of 4.25 out of 5) with sufficient readership (25 ratings or more) to give all users of the application the ability to read the message. Messages in the feed are searchable, though by default they are excluded from the user’s searches.

Users may also add their personal (and private) tags to messages in the feed. Messages in the feed are included in the “Recent Messages” section of the Read All About It home screen.

5.3 Experimental Approach

5.3.1 Surveys and Logs

We required each user to complete an intake survey which focused on demographic information and ascertaining which information sharing methods they currently use, and which ones they like to use. The intake survey mainly consisted of multiple choice and categorical questions, with some Likert-Scale questions at the end to gauge the user’s attitude towards various methods of sharing information; a single open ended question asked for the user’s educational institution, if any. We configured a second survey to be presented no more often than every other week (based on the user’s first use of the application); this survey was intended to examine how they were responding to using the applications to share information and track if and how their preferences changed over time. However, since only three users took the second survey, we are unable to draw any useful data from their answers; no data from that survey is presented here. When it became clear that users were not adopting the applications, the focus of the research changed to examining the reasons for that and the users were asked to take a final survey; thirteen users responded to the final survey over a six week period. The final survey consisted almost entirely of open ended and Likert-Scale questions, with

only three multiple choice questions.

We also instrumented the applications to log page views and actions. This allowed us to see things like the number of the user's friends with open information requests without needing to examine the requests ourselves, or to see which search parameters would become the most popular ones.

5.3.2 Application Users

We chose to recruit our users from six University of California campuses: UC Santa Cruz, UC Davis, UC Riverside, UC Santa Barbara, UC Irvine, and UC Los Angeles. These campuses are all on the quarter system and start classes at roughly the same time. We grouped the campuses into three pairs – UC Santa Cruz and UC Davis, UC Riverside and UC Santa Barbara, UC Irvine and UC Los Angeles – based on overall enrollment [159].

Our initial plan was to simply visit each campus and speak to as many graduate students, postdocs, and professors as we could, while also posting flyers at key gathering points and bulletin boards around the campuses. Later, we changed tactics and began engaging the department staff and asking them to forward a message describing the research to their graduate students, an approach we were able to employ directly at our home campus of UC Santa Cruz. The campuses that relied on face-to-face evangelizing, UC Davis and UC Riverside, had significantly fewer users than the campus recruitment efforts that relied on mailing lists.

Our hope was that by enabling our users to categorize, organize, and search

Tidbitz		Nuggetz		Read All About It	
University	# Users	University	# Users	University	# Users
UC Santa Cruz	29	UC Los Angeles	38	UC Santa Barbara	30
UC Davis	4	UC Riverside	8	UC Irvine	33
Other Universities	2	Other Universities	4	Other Universities	0
Non-University	0	Non-University	2	Non-University	3

Figure 5.9: User Institutions

Tidbitz		Nuggetz		Read All About It	
Gender	# Users	Gender	# Users	Gender	# Users
Female	21	Female	30	Female	40
Male	14	Male	22	Male	26

Figure 5.10: User Genders

their papers and notes in a social environment, we would be able to examine how varying the degree of fun provided by the application affected usage patterns and information sharing. Unfortunately, our total user population across all three applications is under two hundred and most users stopped logging in within two weeks; only three users logged in two weeks after their first time. Therefore, we did not have enough of a response to perform the study we originally envisioned. Six weeks after finishing the marketing phase on the UC Campuses, we took a snapshot of the survey responses. Of the 153 responses, 35 were from Tidbitz users, 52 were from Nuggetz users, and 66 were from Read All About It users. Figure 5.9 provides the breakdown of universities and Figure 5.10 provides the gender breakdown.

5.4 Survey Results

Examining the failure of the applications to capture any sizable user base, and indeed any recurring users at all, begins with the users themselves and an exploration of their current information sharing habits. Then the applications themselves must be examined, to determine if the features were sufficient to meet the needs of the users. Finally, the social context in which the users interacted with the applications is scrutinized to determine how it affected the user’s behavior. In the discussion below, questions from the Intake Survey are labeled with “IS” and questions from the Final Survey are labeled as “FS”.

5.4.1 Information Sharing Habits

Upon logging in for the first time, users were presented with an intake survey. The majority of the questions focused on demographic data, but users were also asked how often they share information using various methods, and how much they liked using those methods. For the purposes of this survey, we defined “information” as news or academic articles, papers, and websites. By doing so, we hoped to focus the user’s attention on work-related information they share, rather than anything they might find on the internet or social information like photos, parties, life events, and so on.

Users were asked to consider the following information sharing methods: email, social networking sites, verbal communication, instant messaging, and written notes. Email is generally assumed to be the most used and popular method of information sharing today. While social networking sites are heavily used to share social information,

it was not clear people would also use them for sharing significant amounts of work-related information. It was not clear how often verbal communication would be used to share information because, while speech is a fundamental part of life, it seems awkward to speak URLs or other internet resources when an emailed link doesn't need to be remembered and is more immediately useful. Instant messaging applications are another social tool that has been popular for many years, but given the requirement for both users to be available at the same time, it was unclear how much use they would see as information sharing mechanisms. Written notes seem almost quaint in the modern world, but we were not sure how often users would resort to them (say, if electronic means weren't available at that moment).

28% of the users reported they share information via email several times per week (IS #9). 22% reported sharing about once per week, and another 30% reported sharing less than once per week. The rest of the users were almost evenly distributed between almost never sharing information via email (8%), sharing once per day (6%), and sharing several times per day (7%). Over three quarters of the users (77%) reported they liked or strongly liked using email to share information, and only 6% said they disliked it; email was the only method that nobody strongly disliked.

Just over half the users (56%) reported they use social networking sites to share information less than once a week or almost never (IS #10). 31% shared information weekly or several times per week. And 14% used social networking sites to share information once or more per day. Almost two thirds (63%) of users reported liking or strongly liking social networking sites for information sharing, but 30% were neutral,

and the rest disliked or strongly disliked them.

Half the users reported using verbal communication, such as in person or over the phone, almost never or less than once a week to share information (IS #13). Another 40% related information verbally on a weekly basis, and 10% on a daily basis. 68% of users reported liking or strongly liking verbal information sharing, and only 11% disliked or strongly disliked it.

Instant Messenger was even less used, with 71% of users reporting they use it to share information less than once a week or almost never (IS #11). 19% of the users utilized IM on a weekly basis, and 10% on a daily basis. People were closely divided in terms of how much they liked using IM to share information, with 34% liking or strongly liking IM vs 28% disliking or strongly disliking it, and 38% neutral.

Paper notes were by far the least popular method of sharing information (IS #12). 84% of users reported they almost never left notes for people, and only 9% reported leaving notes at least once a week. 41% of the users reported disliking or strongly disliking written notes; only 17% reported liking or strongly liking it.

Only 15% of users reported they had a good system in place for managing their notes and papers (FS #14-m). Only 7% reported having a notes and paper management system in place that was too large to change it (FS #14-n); likewise, only 7% reported they would be unlikely to change to a new system with clear benefits for them (FS #14-o). At the same time, some users (15%) reported concerns with trusting their notes to an application that may not be around for very long (as research applications are often short-lived) (FS #14-k), and others (23%) were hesitant to store their notes in an

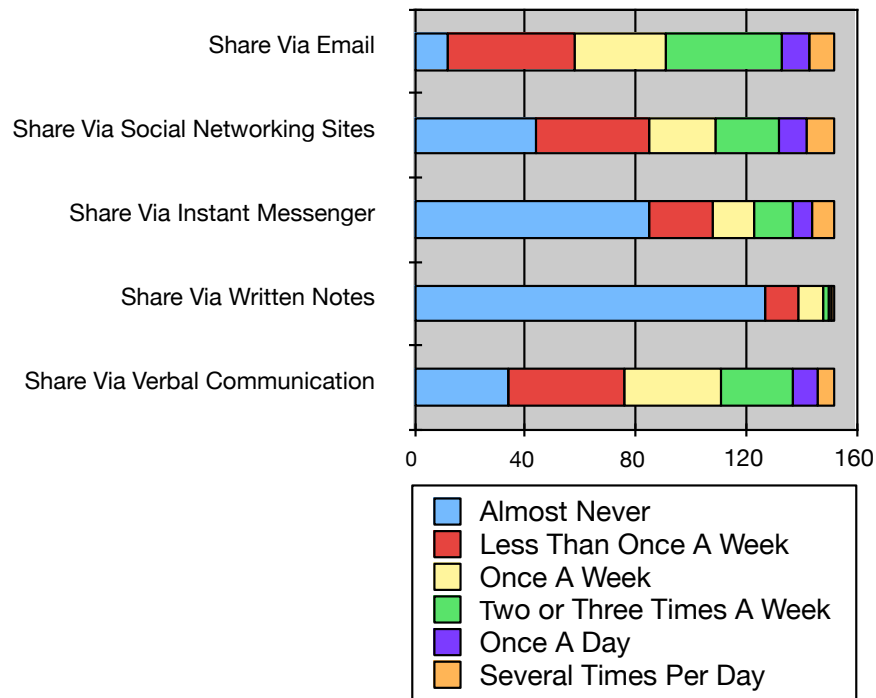


Figure 5.11: Information Sharing Method Use

application they have no control over (these users preferred text files on their computer) (FS #14-l).

5.4.2 Information Sharing Applications

As part of the final survey, users were asked what, if anything, they found confusing about the applications, or if they were easy to use (FS #12). Over half of the respondents (61%) reported very little or no confusion when using the apps, though some (15%) had to come back a few times before they felt comfortable. Fewer than a quarter (23%) reported being confused as to what the application did. When directly asked if they knew how to use the application, only 30% said they did not (FS #14-g).

Users were also asked what features they would expect, or want, to see in a paper/notes organization system (FS #15). Some of the responses were obvious requirements such as “easy to use”, and “reliable”. Others responses included very specialized requirements like a built-in equation editor, and text to speech capabilities. The most requested feature was the ability to cross-reference, or link, notes on one paper with another one; none of the applications supported this feature. The next most requested features were strong organizational tools and keyword searching; as described above, the applications all supported keyword tagging of the papers, along with the ability to search over those keywords.

The rest of the requests included features such as storing and searching the papers themselves (a feature that had to be dropped due to copyright issues), both online and offline access, integration with bibliography tools (EndNote and BibTeX), searching note content, and updating notes. Of these additional features, the applications editing notes as well as searching the content of the notes.

While the applications don’t support every feature the users requested, they do support two of the most requested (organizational tools and keyword searching), note taking and editing, searching notes and other metadata, and obviously the applications have online access. Given this feature set and the user’s requested features, it does not seem unreasonable that some people would find the applications compelling enough to use; additionally, most users reported they understood how the applications benefit them. It also seems unlikely that the addition of cross-references to the feature set would open the floodgates, as none of the survey answers suggested the lack of this

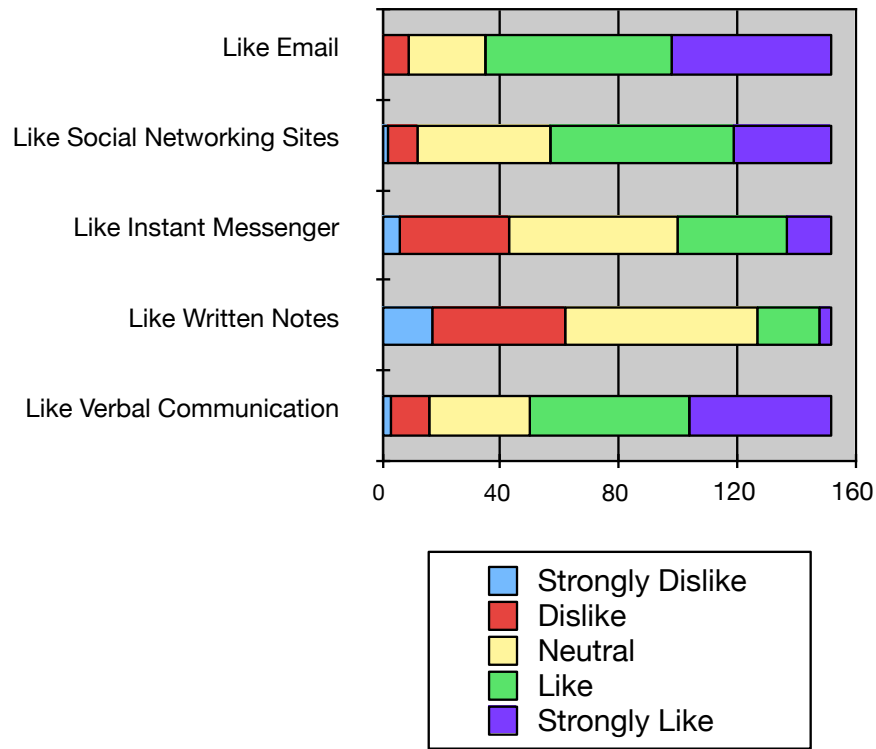


Figure 5.12: Information Sharing Method Satisfaction

feature prevented adoption.

5.4.3 Social Context

The final survey also asked users about the context of the social networking site, Facebook. In terms of experience, the vast majority of users (85%) have been on Facebook for more than a year, most (53%) longer than two (FS #2). All users reported using Facebook daily except under special circumstances (FS #3); at the same time, many (46%) also said they don't have time for Facebook (FS #14-c). All but one user reported an interest in social networking sites, and liking Facebook in particular; the

other user was neutral on both (Final Survey, Questions 14-a and 14-b). Most users even reported they would be willing to use Facebook for work, though about 30% said they would not (FS #14-e). Users were split, however, on whether or not having a Facebook window open while trying to work was inconvenient (FS #14-f). In the end, almost all users said they didn't remember to open the application while they were reading a paper until later on, if they did at all (FS #14-j). When asked under what circumstances they would use the applications more, the 53% of the responses included an expression such as "if more of my friends used it" or "if I were friends with more of my work colleagues" (FS #13). At the same time, some users (23%) also said they don't add professional contacts to their Facebook friends list (FS #13), which makes sense given the overall focus on social life in Facebook; all users said that Facebook is something they do for fun (FS #14-d).

There are are some views that could not be captured by the surveys because only people who signed up to use one of the applications responded to them; users who did not sign up are simply not represented. As part of the recruitment efforts, we spoke with many professors, graduate students, postdocs, and even a few undergraduates on the various UC Campuses. When presented with the research project, almost all professors reported they do not use Facebook at all; those who did have Facebook accounts used them almost exclusively for keeping in touch with family members or close friends and expressed the same hesitation about including graduate students and collaborators in their friends list as the survey respondents. Only a couple of the professors reported they both had Facebook accounts and were friends with their collaborators, students,

and/or colleagues. Additionally, there was a small percentage of graduate students and postdocs who said they do not have, or want, Facebook accounts. For these users, the applications developed here were completely inaccessible, and the “price of entry” (a Facebook account) was simply too high.

5.5 Discussion

There are several lessons to take away from this project, including the relative difficulty in overcoming existing communication methods, and the existence of important sub-groups within the overall group of “academics”.

5.5.1 Information Sharing Habits

Email is clearly king when it comes to sharing information. Its ubiquity, simplicity, and convenience for both the sender and recipient are unmatched by any other method. And while social networking sites are used more often than verbal methods, our users still prefer verbal communication. Instant Messaging shows some promise, but is still rarely utilized, perhaps because the various IM networks don’t talk directly to one another and most don’t allow messages to be sent when the recipient is offline. The applications developed for this project fall under “social networking” because they are a part of the Facebook site. While users reported they did not use social networking sites to share information very often, they clearly like using the sites to do so. This suggests the applications were in a good position to increase the frequency with which users would share information over social networking sites.

5.5.2 Social Groups and Problems Faced

Pinch and Bijker [133] state “A problem is defined as such only when there is a social group for which it constitutes a *problem*,” and further “whether a provisionally defined social group is homogeneous...”. Initially, we considered all academics to be in a single group, and examined the problems they might encounter. All academics spend a significant amount of time examining papers, books, and other resources related to their work, and often keep notes on what they find. Often colleagues will want to share a resource they find with someone else, as happens when a professor sends an article to their graduate student. It is easy to presume that, for these purposes, all academics face the same problems and therefore can be considered a single social group. However, the results presented here suggest that the social group of “academics” is not as homogenous as we hoped and there are sub-groups that must be considered. This is further underscored by Hewitt and Forte’s [83] finding that approximately 33% of students do not believe faculty should be on Facebook, due to “identity management or privacy issues”, and that many people view Facebook as a social venue where it would be inappropriate for students and faculty to connect.

5.5.2.1 Social-Networkers vs. Non

Professors are clear candidates for consideration as their own sub-group. While they might welcome tools to support their work, it is clear that most prefer to keep their work and academic life separate from their social lives, and perhaps want to remain outside the social lives of their students as well. But there are others, who are not

professors, who want the same separation. This suggests two sub-groups that need to be considered: those who are comfortable mixing their work and social lives, and those who prefer a separation. This project may have found greater success if the applications had been built as standalone websites with the option of associating local accounts with a Facebook account via the Facebook Connect API.

5.5.2.2 Electronic Note Takers vs Non

Another sub-division that should be considered is based on where users prefer to read and where they keep their notes. Some users reported a preference for reading from paper and making notes directly on the paper, rather than on the screen; only a few more said their preference was for reading on the screen and keeping electronic notes. Translating written notes to the computer is still a painful process that requires scanning the pages and correcting the OCR process; improving that seems out of the scope of information sharing applications, even though in this context, they would benefit greatly. Almost all users have a mixed approach, depending on the resource in question, and until the vast majority of academic resources are available in digital form (including older publications) it seems unlikely that users would be able to rely entirely on electronic resources. There are features that could be added to better support electronic reading and note taking, such as reading common file formats and allowing users to add notes directly to the file; while that is often supported in the format's native authoring tool, in the context of this work, the information sharing applications would need to support it for several file formats, including PDF and Microsoft Word. Such support may be a

technically challenging feature for a web-based application.

5.5.3 Framing Facebook

Bijker also developed the idea of a “technological frame” [18] which he defined as “the concepts and techniques employed by a community in its problem solving.” For many people, as demonstrated by the professors we spoke with while recruiting users, social networking sites in general do not fall into their technological frame for doing “work” (LinkedIn [119] is something of an exception to this, but its scope is limited to recruiting and job hunting). Facebook in particular, with its development by and for college students, is not generally something one thinks about as a tool for work; many companies even ban employees from using Facebook while on the job [60]. With its historical focus on fun, it is no surprise that users didn’t think to go to our Facebook applications to manage their notes.

Facebook also provides its users with several methods of sharing information already, including a built in IM system, user messaging (similar to email), and wall posts. While these mechanisms are more limited in their feature set than our applications (none of them can be searched, for example), because they are provided by Facebook itself, are featured more prominently than any third party application can be, and don’t require the recipients to have that application installed to see the content of messages.

5.5.4 Application Design

Upon completing the Intake Survey, users were taken to the application's home screen. For a new user, there was nothing there to suggest a next action beyond adding an information request or a bit of information. Providing users with a welcome message or a help document describing the application could have helped drive more adoption. Given the sheer number of notifications, wall posts, event invitations, and fan page announcements a Facebook user is presented with each week, the applications could have used a feature or two that kept users coming back.

5.6 Conclusion

While it is easy to state that work is better when it is turned into play, the experience presented here suggests simply having features that the users want in an fun environment does not guarantee adoption. Especially when operating in a social area (whether a multi-player game, or social networking site) designers must also consider the social ramifications of their decisions. A analysis of the users determined that the major flaw was the failure to recognize the size of the user sub-group that is uninterested in being socially connected to collaborators, even lab mates, within Facebook. Had the applications been developed independently of an existing social networking website, but with the ability to connect to those sites (for the users who wanted to), it is possible they would have seen much greater success. At the same time, given the popularity of email as an information sharing approach, and the number of people who like using it,

there is no guarantee that a social networking tool would capture a significant user base without additional (non-game) features. All in all, our experience provides a cautionary tale of the challenges inherent in turning work into play.

Chapter 6

Information Sharing on Twitter

People share vast amounts of information over social networks every day in the form of status updates, which are generally short textual comments; often these status updates include a link to a news article or other document that provides context for the status update. While most of those status communications are private and shared only with the poster's friends, in a few social networks they are public. Some of the users picked for the analysis described in this chapter are celebrities within their fields. With several fields selected for study, a supplementary model of actual information flow within a large scale social network can be developed, providing some insight into how information might flow through well known figures in an academic discipline compared to the rest of the practitioners — a Stephen Hawking or Ray Kurzweil of a field, compared to the graduate students.

6.1 Introduction

Twitter [157] is a social networking site focused on micro-blogging. Each “tweet” (Twitter parlance for a post) can have at most 140 characters, which forces users to keep their thoughts succinct (or split the thought over multiple posts). By default, all posts are public, and anyone in the world can “follow” (subscribe to) a user’s Twitter feed. However, users are able to make their feeds private and require all followers to be approved. Users can post original content, link to other websites, “retweet” (forward another user’s post), or send public messages directly to other users.

We are interested in examining two types of Twitter users: the “every-person” user and celebrities. The vast majority of Twitter’s users are obviously not celebrities, but the media and bloggers often call attention to entertainers on twitter [4, 81, 90], politicians use it to connect with their constituents [1, 33, 144], businesses use Twitter to connect with customers and attract new ones [8, 142], and leaders and pundits offer advice and commentary on the state of their particular industry [45, 96, 140, 153]. Examining the behavior of celebrities on Twitter and the behavior of regular people could provide insights into the effective spread of information within Twitter and other social networks.

For this study, a celebrity is defined as a Twitter user who is well known in their particular field. Celebrities were identified in each of the following five fields to represent that field: Entertainment, Higher Education, Technology, Business, and Politics. While some of the selected users are familiar names, like Arnold Schwarzenegger or Martha Stewart, others are less well known outside their field. A complete list of the selected

Twitter users for each field can be found in Appendix B. One hundred additional Twitter users were selected at random to compare against the celebrities (the selection criteria is discussed below in Section 6.3.1).

While any post on Twitter can be considered information sharing, not all posts are equal. This study focuses on three types of tweets that suggest the poster has read something on the internet, and felt that it was interesting enough to share with the people who follow them: tweets containing URLs, retweets without URLs, and retweets with URLs. A distinction is made between retweets with and without URLs because posts with URLs suggest that two users looked at the URL in sequence — the user who posted the original tweet and the user who retweeted it, as opposed to retweets without URLs which only requires the follower to read the original tweet as pass it on; the URL in this case acts as the original information source — and each decided to share it with their followers, potentially indicating a different quality or style of information.

This study examines the difference between tweets posted by celebrities and those posted by a random sample of Twitter users (henceforth referred to as “regular users”), the type of information they share, and factors that affect the spread of information via retweets. It is easy to assume celebrities know of their status in their field and are on Twitter to connect with people interested in their work (fans, if you will), while the average person is more interested in sharing information about their lives with friends and family. Celebrities often use their clout as a soapbox, assuming their fans will listen to what they say and potentially pass the message on to others, a behavior one would expect to see in Twitter.

In addition to the differences in behavior between celebrities and regular users, the study also addresses questions about re-tweetability (a topic covered in more detail by Zarrella in his self-published study [177]). The basic premises proposed here are that more opportunities for a tweet to be read will increase its chances of being retweeted, and tweeting more often will increase the number of a user's tweets that are retweeted. Twitter presents tweets to the user in a linear fashion, with the most recent tweets at the top. This means that to see tweets from previous days or weeks, a user must scroll to the bottom of the page and request more tweet content; the farther back one wants to read, the more time it takes to load the content and most users will not bother to load more than a few pages [88]. That suggests that tweets posted during hours or days in which people are likely to be using a computer (day time versus night time, for example) are more likely to be retweeted. Finally, if a user tweets more often, they will be more likely to have a post on the first page of their follower's content (analogous to being "above the fold" in a newspaper), and that could lead to a greater percentage of their tweets being read and retweeted by their followers.

The hypotheses explored in this study are:

1. *Celebrities post more original content and retweet less often than regular users.*

Celebrities might be more likely than regular users to use Twitter as a soapbox to share their thoughts with interested people (their followers), and spend less time reading tweets from other users.

2. *Celebrity tweets are retweeted more often than the tweets of regular users.*

Just looking at a supermarket news rack, one can tell there is a strong interest in

what celebrities have to say. Followers are likely to demonstrate that interest by retweeting what celebrities say.

3. *Information shared by celebrities, in both links to external sites and retweets, is more likely to be related to their work and general community than anything else.*

Here as well, celebrities might be more likely to keep their content focused on their field than regular users would.

4. *Retweets are most likely to be posted during times of heavy Twitter usage.*

One has to be using Twitter in order to retweet; the time of day when most tweets are posted is probably also the time of day most retweets are posted.

5. *Retweets are most likely to be posted during days of heavy Twitter usage.*

This expands the previous hypothesis to the day of the week; whichever day of the week sees the most tweets posted is also likely to see the most retweets posted.

6. *Users who tweet more often, whether celebrity or regular user, will be retweeted more often.*

A user who tweets more often is more likely to have their tweets at or near the top of their follower's Twitter feeds, making it more likely that a follower will see at least one of their tweets and retweet it.

7. *Users with more followers, whether celebrity or regular user, will be retweeted more often.*

Similarly, having more followers will also increase the chance that at least one follower will read and retweet a user's tweet.

This chapter is organized as follows. Previous work related to knowledge sharing and Twitter is discussed in Section 6.2. The experimental design for this study, including criteria for selecting users and the data collected from those users, and the analysis is described in Section 6.3. The results of the analyses are described in Section 6.4, followed by a discussion of the findings in Section 6.5 before concluding in Section 6.7.

6.2 Related Work

Knowledge transfer has been studied widely before Twitter was invented. Wu et al. [170] compared information flow in social groups by examining their email history and found that information spreads much like an epidemic, except that it is more likely to be limited than a disease as interest in the information tends to degrade sharply once it leaves a particular social group.

Since its launch in 2006, Twitter has been a rich source of content (rich enough that the Library of Congress will be archiving the entirety of Twitter’s public content [156]), inspiring research that spans from predicting a movie’s success during its opening weekend [166] to comparisons between a user’s tweets and entries in more traditional diaries [89]. Brooks and Churchill discuss how Twitter has changed the way some people monitor local news and events, or use their followers as a private recommendation engine [25]. Subramanian and March find that Twitter users often make an effort to post interesting information in their tweets, lest they seem boring to their followers [149].

As the primary mechanism of knowledge transfer on Twitter, retweeting has been examined in a myriad of ways. Boyd et al. [22] explored the structure of retweets across Twitter, describing several different formatting styles and exploring how information and its context changes through the process of being retweeted. Lerman and Ghosh [115] compared the spread of news on both Digg and Twitter, finding that Twitter’s younger, and sparser, network did not propagate information as well as Digg’s more established network. In both Digg and Twitter, most “votes” (retweets on Twitter) occur within a short time of the original tweet being posted.

Suh et al. [151] presented a detailed exploration of factors that affect the retweet-ability of a particular tweet. They found positive correlations between retweets and the number of followers a user has, the number of people that user follows, the inclusion of a URL or hashtag, as well as the age of the account. However, Cha et al. [34] found that the number of followers did not have a strong influence on the number of retweets. Rather, their findings indicate that users who are often mentioned by others are also often retweeted, and vice versa, but the number of followers did not have a strong influence on either retweets or mentions. The discrepancy between Suh et al. and Cha et al.’s findings could stem from their different methodologies. Suh et al. normalized their data but Che et al. found that “normalization failed to rank users with the highest sheer number of retweets as influential”; Che et al. state that normalization might have led to different results. Additionally, Suh et al.’s findings of a URL in 28.4% of retweets differs strongly from a previous study by Dan Zarrella [177] which found 56.69% of retweets contain a URL. Zarella’s study also looked at how retweets

related to many variables including speech and grammar patterns, time of day, day of the week; he found that better grammar produces more retweets, as did tweets posted in the afternoon or evening rather than the morning, and tweets posted on Mondays, Fridays, and Saturdays.

Romero et al. [141] also examined user influence in Twitter and developed an algorithm to predict it that outperforms simple number of followers, or even more advanced methods such as PageRank. Their algorithm adds a measure of user passivity to the calculations, to account for the relatively low level of retweets found in their dataset. The results of their study again show that well known Twitter users with large numbers of followers do not necessarily have more influence.

The lack of correlation between number of followers and influence was also found by Ye and Wu [175]. In their study, which focused on how breaking news spreads through Twitter, they found that influence is better measured by the number of replies a user has, followed closely by the number of unique users replying to them, the number of unique users retweeting them, and the number of times their messages are retweeted.

Khrabrov and Cybenko [101] used Twitter’s Streaming API to “drink from the firehose”, collecting 100 million tweets over the course of a month (though still a sub-set of all tweets posted during that time). They then created a graph in which Twitter users were nodes and edges were created from the poster to any users who they mentioned or replied to in their tweets using `@username`. Each day they computed a PageRank for each user, and sort the users by their ranking to create their *dirank*. They focused on users who maintained or increased their *dirank* over the course of the study

(a single instance of a lower rank was enough to filter the user), and those who had accelerated activity over the course of the study. The users identified by their work included journalists, pop stars, actors, and their fans. They found that these accounts, especially those created by fans, often engaged in activities designed to increase their visibility within the Twitter network, such as offering to @mention other users who @mentioned them, or creating multiple accounts.

Wu et al. [171] examined the relationship between celebrity Twitter accounts and non-celebrity users. Celebrity accounts were discovered by using the List feature of Twitter to find accounts that frequently appeared in lists; they were classified semi-automatically based on the keywords associated with the lists into Celebrities (in the Hollywood sense), News Media, Organizations, and Blogs. Their work examines the network in terms of these users types and describes who they are likely to listen to, who is likely to follow them, and what kind of news they are interested in. They also find that a relatively small number of users account for half the links found in tweets.

Huberman et al. [87] examined the strength of the connections between Twitter users. They defined a user's "friends" as those people the user had directed two or more posts to. The number of friends a user has was a better predictor for the level of that user's activity than the number of followers, with more friends suggesting more posts by the user. Additionally, they found that the number of friends in their data set became saturated around 30 or 40, while the number of followers continued to increase. They suggest that information is more likely to flow from the user to their friends than it is to flow from the user to the average follower.

Influence and information diffusion within Twitter was also examined by Kwak et al. [107]. Unlike the work presented here, they focused on “trending topics”, or tweets that include a hash-tag that is popular at the moment; for example, when a celebrity such as Michael Jackson passes away, many of the trending topics will reference that celebrity in some way. They found that trending topic tweets reached an average of 1000 users, even if the initial audience was small, and that fewer than five users separate the initial poster from most of the eventual recipients. Additionally, they found that the median time between a tweet and its retweet was just under an hour, and 75% of all the retweets happen in the first 24 hours. However, by focusing on trending topics, Kwak et al. ignored the large swath of tweets that are not of interest to the general population.

André et al. also examined tweet content in an effort to determine the value of that content [6]. They found that a quarter of tweets were considered not worth reading by followers, and only a little more than a third of tweets were considered worth reading; followers were neutral on the rest. Tweets considered not worth reading were often boring to the followers, or lacked the context needed for the followers to understand the tweet. Tweets that consisted mostly of hash-tags and @ mentions were also rated as not worth reading. In contrast, tweets that informed or were humorous were more likely to be considered worth reading by followers.

Several of the works cited here have examined their data set to understand how the number of followers a user has and might affect how often that user is retweeted and the user’s influence; this work also looks for correlations between followers and retweets, but does not try to measure influence. This study contrasts with the other

studies discussed here in that they included organizations and corporate entities in their categories; as we discuss in Section 6.3, our study sought to focus on individual Twitter users. Additionally, this work compares the behavior and content of regular people on Twitter with that of known celebrity accounts, rather than trying to identify each group automatically.

6.3 Experimental Design

To address the hypotheses about user behavior and the type of information being shared on Twitter, rather than building a large database with millions of tweets across hundreds of thousands of users, the approach here selected a relatively small number of users, half of whom are celebrities selected in an ad hoc manner and half of whom were selected at random. This approach allows comparison between the relative influence between celebrities and the regular Twitter user. Once those users were identified, an automated script was used to harvest their tweets, and other data about them.

6.3.1 Twitter User Selection

The tweets focused on in this study are from two different types of users: celebrity users and regular users. Celebrity users are defined as people who are well known in a particular field, even if they are not household names. Regular users were selected at random from the range of possible Twitter User IDs.

6.3.1.1 Celebrity Users

Celebrity users came from five fields: Technology, Politics, Business, Entertainment, and Higher Education. The celebrity users were hand-selected haphazardly¹ based on their regular use of Twitter, with an emphasis on finding personalities who use Twitter directly, rather than through an assistant (though there is no method of determining this with any degree of certainty). Some celebrity users were found by simply performing a search of their name and “Twitter”, while others were found via websites dedicated to tracking Twitter accounts of people of interest in a particular field; for example, `tweetcongress.org` tracks members of the US Congress who maintain Twitter accounts.

6.3.1.2 Regular Users

The regular users were drawn from a pool of randomly selected Twitter User IDs. The upper bound of possible User IDs was determined by creating a new Twitter account and using its ID, assuming that Twitter assigns User IDs sequentially; each random user ID was then examined to ensure they met certain criteria (described in Table 6.1). 5,369 Twitter accounts were examined in selecting the 100 regular users. Table 6.1 describes the criteria and the number of users rejected because they did not meet it. The examination process involved a mixture of automated and manual processes. Criteria such as the “Existence” and “Usage” are easily determined by a com-

¹Haphazard selection means the users were selected based on their accessibility (how easy it was to find them). It does not guarantee that the selected users are representative of the whole. Standing on a street corner or outside a store to survey passers-by is the most well known form of haphazard sampling.

Table 6.1: Regular User Selection Criteria And Rejection Counts

Name	Description	# Rejected
Existence	A user with that user ID must exist.	959
Public	The user's tweets must be public.	93
Language	The user's tweets must be predominantly in English so that we can describe and categorize the content.	145
Not Spam	The account must not appear to represent spam.	23
Personal	The user must not represent an organization's official Twitter feed, whether it be a corporation, community group, or governmental agency.	39
Longevity	The user's account must be at least six months old.	243
Usage	The user must have tweeted at least twenty five times and not have breaks in usage spanning more than a month.	3769
Recency	The user must have tweeted within the last month.	98

puter, while criteria “Not Spam” and “Language” must be determined manually; the manual process was performed by a single person. Only one criteria for rejection was recorded, even in cases where a user met several of the them. Corporate and organizational accounts were screened out because by their very nature, they can be managed by multiple people; this makes comparisons to accounts managed by a single person difficult. The Longevity criteria helped ensure that all the regular users in the study were active Twitter users with accounts that had time to build their list of followers. The Usage and Recency criteria were used to make sure the accounts had not been abandoned, and potentially losing followers because of that, and were utilizing Twitter as both content generators and content consumers.

Both the Language and Usage criteria likely introduced some amount of sample bias into the results. The Language criteria was necessary to perform the content analysis without use of translators. The Usage criteria was needed because the hypotheses put forth are about tweet content, and if a user is not producing content, their data does not contribute to the analysis. Of the users rejected for not meeting the Usage criteria, over half had no tweets at all. This means the regular users examined here are not representative of all twitter accounts, which include a significant number of users who do not produce any content, even by retweeting the content of others.

6.3.2 Data Harvesting

Twitter’s API [158] provides third parties nearly complete access to a user’s data and tweets, assuming the Twitter account used to make the requests are made has

Table 6.2: Twitter User Data

Data	User Reported
User ID	No
Screen Name	No
Real Name	Yes
Location	Yes
Time Zone	Yes
Description	Yes
URL	Yes
Profile Image URL	No
Number of Followers	No
Number of Friends	No
Number of Tweets	No
Account Creation Date	No
Number of Favorite Tweets	No

Table 6.3: Twitter Tweet Data

Data
Tweet ID
Tweet Creation Date
Content
Author’s User ID
Retweet Flag

access to that user’s data. The most significant limitations are that only the most recent 3,000 tweets can be accessed for a particular user, and only 100 retweets can be accessed for a particular tweet. 50 users in our study (43 celebrities, and 7 regular users) had exceeded this threshold at the time their data was collected, so not all of their tweets were accessible. It was not clear how many tweets from the study were retweeted more than 100 times, as this count was not available in the dataset.

We created a script to harvest data from Twitter, with an accompanying website to monitor the process. Over the course of several weeks, the script gathered data about the users and their tweets. The data Twitter provides for each user is listed in Table 6.2 and the data provided for each tweet is listed in Table 6.3. Some data about the users is self-reported by the user when they create their account and was not considered reliable enough to use for analyses; these fields have been indicated in the table.

Also of note is the “Retweet Flag” data item in Table 6.3. While Twitter has an official “Retweet” action in their user interface that will automatically flag the new

tweet as a retweet, there are situations where retweets are not flagged. Users might copy a tweet and paste it into the new tweet field, following the convention of adding the characters “RT” to indicate they are retweeting someone else, and some third party applications may not correctly indicate that the tweet they are submitting is a retweet. In addition to using “RT”, Twitter users will often use the “via @user” convention to attribute some or all of their tweet’s content to another user; this is commonly seen when the tweet contains a link from another user’s tweet with original text from the attributing user. To account for both situations, all tweets containing the characters “RT” or “via” were manually scanned to find any retweets or attributed tweets that had not been automatically flagged as such. Manual checking was required to prevent the inclusion of tweets where “RT” was used as abbreviation, requests to “please RT this” as a method of propagating the content, and uses of the word “via” that did not follow the “via @user” convention.

When accessing a user’s timeline, the data returned for each tweet does not indicate whether or not it has been retweeted. To make this determination, after the script finished gathering all the basic user and tweet data, it requested all the retweets (subject to the Twitter API’s 100 retweet limit) for every tweet collected.

6.3.3 Tweets and Retweets and Retweeted Retweets and Links (Oh my!)

Anything posted on Twitter is considered a tweet. This study breaks those tweets into various categories including plain tweets, retweets, tweets with links, retweets

with links, and so on. Here are some examples of the types of tweets discussed.

A **plain tweet** from Mayor Cory Booker:

Am I eligible to try out?

A **retweet** from the TV show “The X-Factor USA”:

Yes! We know you have The X Factor. RT @CoryBooker: Am I eligible to try out?

A **retweeted retweet** from Mayor Cory Booker

*Yes, my singing is X-tremely X-cruciating RT @xfactorxtra Yes! We know you
have The X Factor. RT @CoryBooker: Am I eligible to try out?*

In these examples, the tweets are being exchanged between two Twitter accounts as part of a conversation, but this study makes no distinction between exchanges such as this and tweets that are passed between three different users; in either case, the third response is a retweeted retweet. The study does distinguish between tweets without links to external websites, as seen here, and those that include links. Had a web link been included in the original tweet (perhaps to a call for auditions in Mayor Booker’s city), these would have been termed a **tweet with link**, **retweet with link**,

and **retweeted retweet with link** respectively.

6.3.4 Statistical Analysis

The following sections describe each of the analyses done on the data collected. Section 6.3.4.1 describes the tests used in the subsequent sections. Sections 6.3.4.2 and 6.3.4.3 discuss the Tweet Behavior of the users in this study and their followers, respectively. The tweet content analysis is described in Section 6.3.4.5. Finally, we look for some effects that increase the likelihood of a tweet being retweeted, including the time and day the tweet was posted in Section 6.3.4.6 and the frequency of the user's tweets in Section 6.3.4.7. All analyses were performed using R 2.10.1 running within the 64-bit R.app GUI for MacOS X build 1.31.

6.3.4.1 Statistical Tests

For the analyses in Sections 6.3.4.2 and 6.3.4.3, the Shapiro-Wilk test [143] is used to determine if the data comes from a normal distribution. The data used for these analyses are percentages of observed counts; since none of the data gathered in this study is normally distributed, non-parametric tests are used for the analyses. The Kruskal-Wallis one-way analysis of variance [106] is used to determine if data for each community comes from the same distribution, however this test only determines if all the communities are the same or not. When the Kruskal-Wallis test finds differences between communities, pairwise Wilcoxon signed-rank tests [169] are used to determine which specific data sets are different and which are the same.

When analyzing the tweet content in Section 6.3.4.5, the data consists of observed counts, rather than percentages, so Pearson’s chi-squared test [132] is used to determine if the distributions of observed content types are the same.

The Kolmogorov-Smirnov test [124] is used in Section 6.3.4.6 on data that is ordered (time-based) to test whether or not behavior depends on the specific order.

In Section 6.3.4.7, generalized linear models [125] are used to perform linear regressions and determine how the test variables (tweet rate and number of followers) affect the percentage of a user’s tweets that are subsequently retweeted by one or more of their followers; a Poisson distribution was used as the canonical link for the model. The generalized linear models are then used to predict continuous responses (retweet percentages) along a full range of inputs (tweet rates or number of followers) using a least-squares fit; when plotted, these predicted responses allow trends to be visually identified.

6.3.4.2 Tweet Behavior of the users

We assessed the differences among six groups of Twitter users (five groups of celebrities and a group of randomly selected users) with respect to three types of knowledge transfer tweets: retweets (without a link), tweet with a link, and retweets with a link. The metrics used for this analysis were the percentage of each of the of those three tweet types with respect to all of the user’s tweets.

$$\%RT_{user} = \#retweets_{user} / \#tweets_{user} \quad (6.1)$$

$$\%TwL_{user} = \#tweetsWithLinks_{user} / \#tweets_{user} \quad (6.2)$$

$$\%RTwL_{user} = \#retweetsWithLinks_{user} / \#tweets_{user} \quad (6.3)$$

Normality of the percentages was assessed both visually and using the Shapiro-Wilk test; non-parametric tests were chosen because data violated assumptions of parametric analyses. Kruskal-Wallis tests were performed on each tweet type with respect to the community to determine which, if any, tweet behaviors differed between communities. Finally, pairwise Wilcoxon signed-rank tests were used to compare each community with the others for each type of tweet behavior.

6.3.4.3 Retweet Behavior of the followers

The process of examining the difference in retweet behavior between the different types of users is nearly the same (the same groups of users were used as above). However, in this case, the actions of the user's followers are the subject of the analysis, as the followers are the ones performing the retweet. This allows the addition of the user's regular tweets in the analysis as a distinct group, comparing follower behavior between tweets that contain links and those that don't (regardless of whether the tweet was a retweet by the user), and the creation of an overall picture across all of the user's tweets. Given these groups, the metrics were computed, as in Section 6.3.4.2, as a percentage of the total for all the user's collected tweets.

$$\%RTed\ T_{user} = \# \textit{retweeted plainTweets}_{user} / \# \textit{tweets}_{user} \quad (6.4)$$

$$\%RTed\ RT_{user} = \# \textit{retweeted retweets}_{user} / \# \textit{tweets}_{user} \quad (6.5)$$

$$\%RTed\ TwL_{user} = \# \textit{retweeted tweetsWithLinks}_{user} / \# \textit{tweets}_{user} \quad (6.6)$$

$$\%RTed\ RTwL_{user} = \# \textit{retweeted retweetsWithLinks}_{user} / \# \textit{tweets}_{user} \quad (6.7)$$

$$\%RTed\ OT_{user} = \# \textit{retweeted overallTweets}_{user} / \# \textit{tweets}_{user} \quad (6.8)$$

Note that Equation 6.8 uses all retweeted Tweets, regardless of type; its result will be the sum of the results of Equation 6.4 through Equation 6.7.

Normality of the percentages was assessed using the Shapiro-Wilk test; non-parametric tests were chosen because data violated assumptions of parametric analyses. Kruskal-Wallis tests were performed on each tweet type with respect to the community to determine which, if any, retweet behaviors differed between followers of the users in the different communities. Finally, pairwise Wilcoxon signed-rank tests to compare each community with the others for each type of tweet behavior.

Table 6.4: Some fake Tweet data

	User's Tweets	Follower's Retweets
Plain Tweets	500	250 (Equation 6.4)
Retweets	250 (Equation 6.1)	125 (Equation 6.5)
Tweets w/ Links	200 (Equation 6.2)	100 (Equation 6.6)
Retweets w/ Links	50 (Equation 6.3)	25 (Equation 6.7)
Total Tweets	1000	500 (Equation 6.8)

6.3.4.4 Behavior Computation Examples

Table 6.4 presents a contrived data set for a single user. In this data set, a user has posted 1000 tweets, 100 of which have been retweeted by that user's followers. For simplicity, this data set has a 50% retweet rate, regardless of the type of tweet. When a cell is used by one of the equations above, the equation is noted; the "Total Tweets" values are used by all the equations listed in the same column. The first column is used in equations described in Section 6.3.4.2 and are used to describe the behavior of the user; the results of the equations are shown in Table 6.5. The second column is used in equations described in Section 6.3.4.3 and are used to describe the behavior of the user's followers; the results of the equations are shown in Table 6.6.

6.3.4.5 Tweet Content

To analyze tweet content, a random sample of 1,000 retweets, 1,000 tweets with links, and 1,000 retweets with links was taken from the collected tweets (3,000 unique

Table 6.5: User behavior in fake Tweet data

	User's Tweets	Results
Plain Tweets	500	(Not used)
Retweets	250 (Equation 6.1)	$\%RT_{user} = \frac{250}{1000} = 25\%$
Tweets w/ Links	200 (Equation 6.2)	$\%TwL_{user} = \frac{200}{1000} = 20\%$
Retweets w/ Links	50 (Equation 6.3)	$\%RTwL_{user} = \frac{50}{1000} = 5\%$

Table 6.6: Follower behavior in fake Tweet data

	Follower's Retweets	Results
Plain Tweets	250 (Equation 6.4)	$\%RTed T_{user} = \frac{250}{1000} = 25\%$
Retweets	125 (Equation 6.5)	$\%RTed RT_{user} = \frac{125}{1000} = 12.5\%$
Tweets w/ Links	100 (Equation 6.6)	$\%RTed TwL_{user} = \frac{100}{1000} = 10\%$
Retweets w/ Links	25 (Equation 6.7)	$\%RTed RTwL_{user} = \frac{25}{1000} = 2.5\%$
Total Tweets	500 (Equation 6.8)	$\%RTed OT_{user} = \frac{500}{1000} = 50\%$

tweets in total). In each group of 1,000 tweets, 500 came from regular users, and 100 from each of the five celebrity communities. Five people were asked to examine 1,200 of 3,000 tweets, so that each tweet was examined by two different people. The examiners were tasked with categorizing the content of the retweeted portion of the tweet or the content of the link in the tweet into one of nineteen categories. The categories were pre-selected by a single person reading several hundred tweets; these initial categorizations were then discarded and not used in the analysis. The examiners could use one of the pre-selected categories or suggest new ones; no new categories were suggested.

Detecting differences using all nineteen categories required a much larger sample size than the examiners would have been able to categorize in a reasonable time, so the original nineteen categories were combined to make four super-categories, “Commentary”, “News”, “Personal”, and “Other”. Figure 6.7 shows the four super-categories and the original nineteen categories. The super-categories are defined as:

Commentary Content that expresses an opinion or is part of a discussion.

“much appreciated! RT @AnaiRhoadsorg: Love Craigslist? Follow the founder - official page - @craignewmark”

News Content that is traditional news.

“@SenateBanker: Dodd says on floor he will co-sponsor Sanders Amndmnt with a change to protect Fed independence.”

Personal Content related to someone’s personal life such as family photos or videos, or posts that promote the person’s own work.

“I posted a new vid. YouTube .com/FranDrescherSong i need another day or two for the surprise. Goin 2c art & hear music 2nite w th BF”

Other Content that could not be categorized, usually due to an error.

“Ah! lol ... <http://lnk.ms/3xxgg>”

The Commentary category represents opinions, suggestions, and discussions that Twitter users might have, including one time question and response exchanges. This is distinct from the News category, which contains content that would be generally considered as fact (even if that fact is later disputed, as happens when pranksters report the untimely death of a celebrity). The Personal category includes any content that relates to the personal life of the content’s creator, whether pictures from their life, or announcements of a new project, or questions posed to other Twitter users. The Other category represents content that is unreadable, has been corrupted somehow, or is no longer available, and therefore cannot be categorized. These four categories (especially the first three) represent three distinct types of tweets; one would expect the mix of tweets to be different for different types of users.

All tweets in which the two examiners disagreed on the overall category were removed. For example, if one examiner classified a tweet as “Business News” and the other classified it as “Technology News”, this was counted as agreement under the “News” category. However, if the first examiner classified a tweet as “Business News” and the other classified it as “Self Promotion”, this was counted as a disagreement and that tweet was not included in the rest of the analysis.

After narrowing the dataset to only tweets with content categories that had

Table 6.7: Tweet Content Categories

Commentary	News	Other	Personal
Advice	Business News	Humor	Personal Pictures
Commentary	Entertainment News	Unknown	Personal Videos
Public Reply	General News		Self Promotion
Quote or Saying	Higher Education News		Announcement
Review	Sports News		Question or Solicitation
	Technology News		
	Political News		

agreement between the examiners, a Pearson's Chi-Squared Test was performed on two views of the data to determine if the distribution of the observed categories differed between the user types in the view. The first view contained both celebrities and regular users and compared the observed categories for those two groups of users. The second view contained only the celebrity users, and compared the observed categories for five communities of celebrities. Because some tweet categories had low levels of observation, further breakdown of the data (by tweet type, as in the analyses in Sections 6.3.4.2 and 6.3.4.3) could not yield statistically valid results.

6.3.4.6 Tweet Timing

For the analysis of tweet timing (time of the day and day of the week), the tweets from the celebrities and regular users, but not their followers, were placed into

two groups, one based on the day of the week and the other based on the hour at the time of the post. Each vector of counts was tested with pairwise Kolmogorov-Smirnov tests to determine if their distributions were the same. Since the types of tweets occur at different rates, a second set of plots is needed to compare the shape of the resulting curve between the various types of tweets; for these plots, a scale factor is computed for each tweet type (see Equation 6.9).

$$scale = 100 / \max(\text{tweetCountVector}) \quad (6.9)$$

Then, the counts and the scaled values were each plotted for visual inspection, including three separate plots for: all users, celebrity users only, and regular users only; six plots in total.

6.3.4.7 Tweet Rate and Followers

A user's tweet rate can be represented as tweets per day; this is calculated by dividing the number of tweets collected from the user by the number of days between the first collected tweet and the last collected tweet.

$$tweet\ rate = \#\ tweets_{user} / (date_{lastCollectedTweet} - date_{firstCollectedTweet}) \quad (6.10)$$

Three generalized linear models (GLMs) were constructed to test if the percentage of a user's tweets that were retweeted by their followers depended on the user's overall tweet rate, and another three GLMs were constructed to see if the percentage of a user's retweeted tweets depended on the number of followers; all GLMs used linking

poisson distributions due to non-normality (see Figures 6.18 and 6.20). Percentages of retweeted tweets were rounded to the nearest whole number to satisfy the requirements of Poisson distribution modeling. The outputs from the models were then used to calculate the best-fit relationship between the two variables in each model.

6.4 Results

This section presents the results in the same order as the analyses were described in Section 6.3.4. The tweet behavior of the users in the study is examined in Section 6.4.1 and the behavior of their followers is examined in Section 6.4.2. Section 6.4.3 examines the findings of the tweet content analysis. The effects of the time of the day and the day of the week on being retweeted is described in Section 6.4.4, while the effects of the frequency of tweets, and the number of followers on being retweeted is described in Section 6.4.5.

Note that for all the boxplots presented in this section, there are data points for each community of celebrity users, and the “Celeb” data point includes all celebrity users for that analysis.

6.4.1 Tweet Behavior of the users

Shapiro-Wilk tests rejected the null hypothesis that the data comes from a normal distribution for each type of tweet, shown in Table 6.8. Figure 6.1 shows the breakdown of the types of tweets that were harvested broken down by the groups of users.

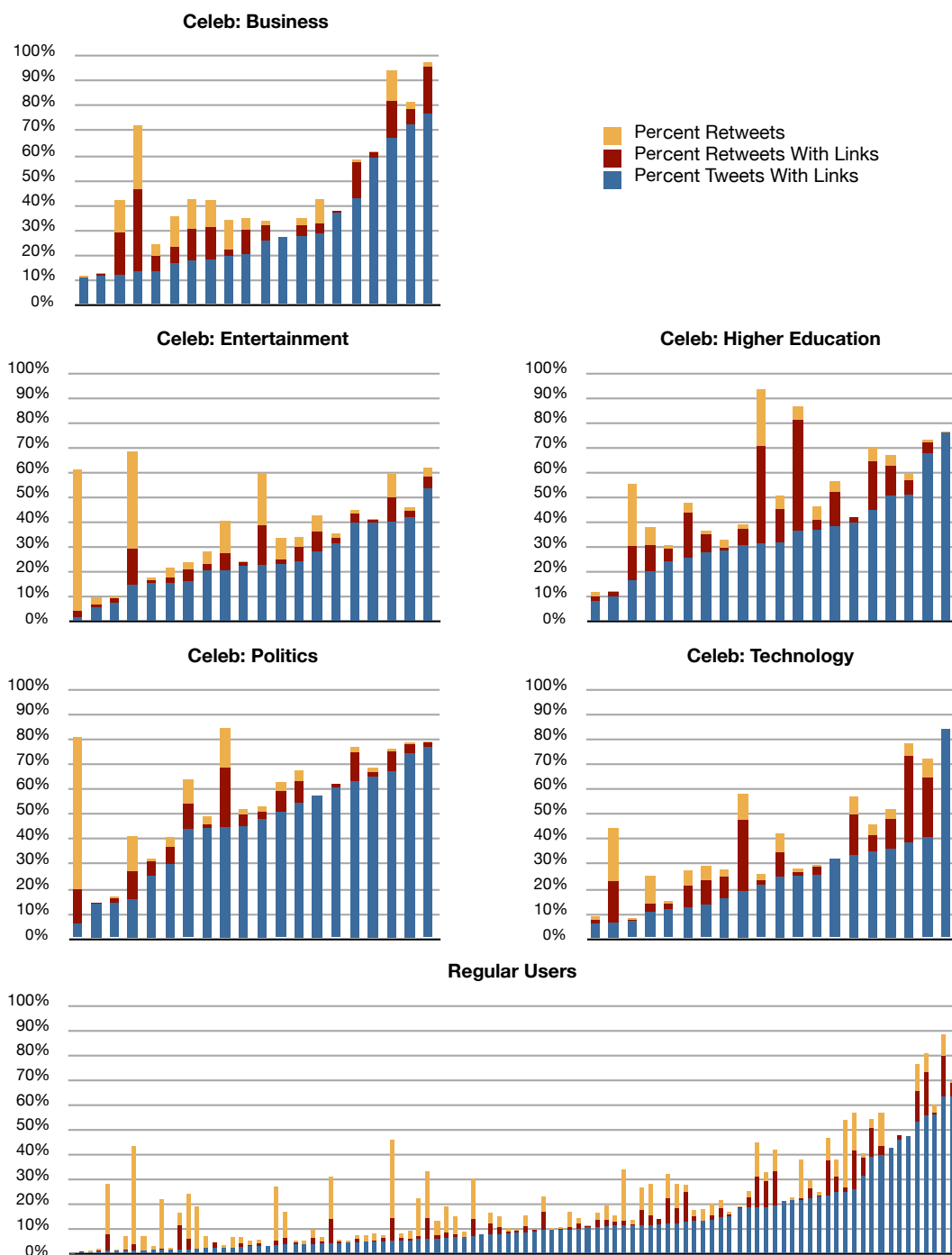


Figure 6.1: Tweet Distributions for each category of user. Each user is represented by a bar along the x-axis, sorted by Percent Tweets With Links. The space between the top of each bar and 100% represents the user's plain tweets.

Table 6.8: Shapiro-Wilk Analysis of Tweet Types

Tweet Type	W	p-value (\leq)
Retweets (RT_{user})	0.6381	2.2×10^{-16}
Tweets with Links (TwL_{user})	0.9107	2.346×10^{-12}
Retweets with Links ($RTwL_{user}$)	0.76	2.2×10^{-16}

6.4.1.1 Tweets with links (Equation 6.2)

The median value for the percentage of celebrity tweets with links (TwL_{user}) was 27.350%, and the maximum value was 83.650%, compared to a median of 8.211% and maximum of 63.158% for regular users. All users in the study posted at least one tweet that contained a link.

Figure 6.2 shows boxplots for the percentage of tweets with links for the sampled users. The Kruskal-Wallis test results for the percentage of tweets with links suggested differences in the groups (chi-squared = 93.3689, $p < 2.2 \times 10^{-16}$). The pairwise Wilcoxon signed-rank test confirms what visual inspection of the boxplot in Figure 6.2 suggests, regular users are less likely to post tweets with links than all groups of celebrities ($p < 0.003$). Wilcoxon tests also show that politicians posted more tweets with links than entertainers and technologists $p < 0.031$, but all other comparisons between celebrities rejected the null ($p > 0.902$).

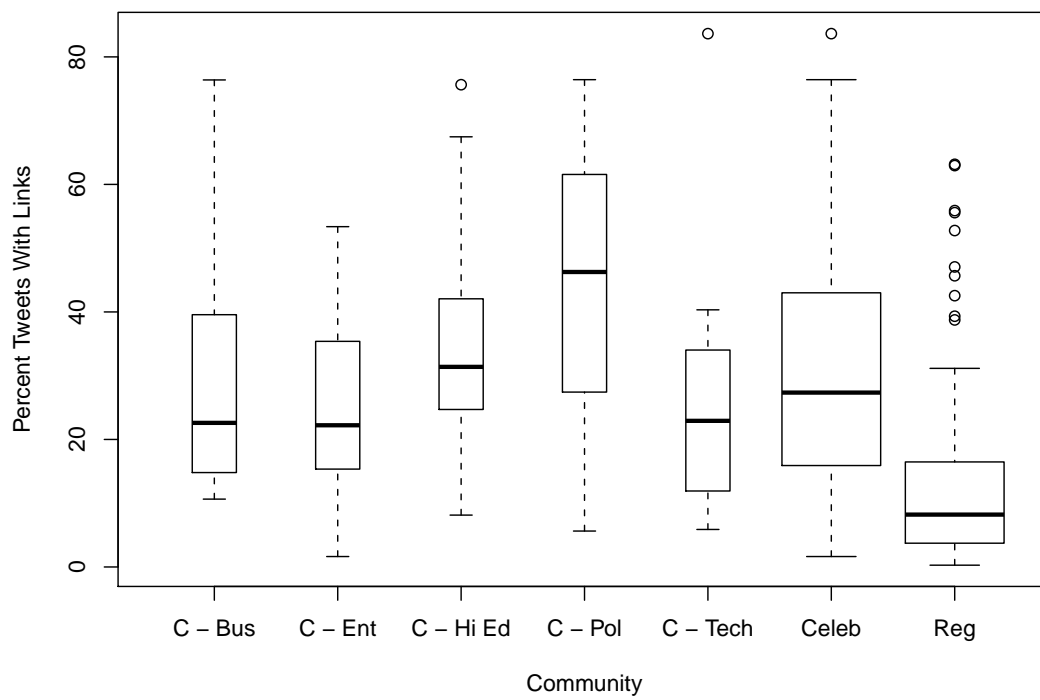


Figure 6.2: Percent Tweets With Links Boxplot

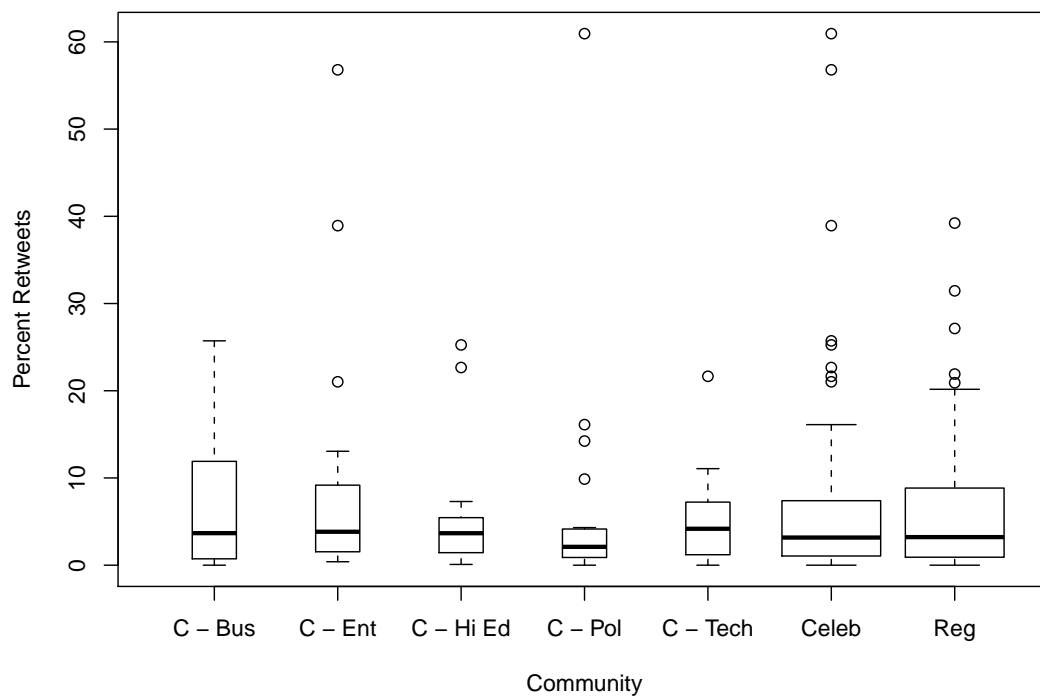


Figure 6.3: Percent Retweets Boxplot

6.4.1.2 Retweets (Equation 6.1)

The median value for the percentage of celebrity retweets (RT_{user}) was 3.169%, and the maximum value was 60.946%, compared to a median of 3.218% and maximum of 39.225% for regular users. While most user groups had at least one user who never retweeted, every user in the Entertainment and Higher Education communities retweeted at least once.

Figure 6.3 shows boxplots for the percentage of retweets for the sampled users. The Kruskal-Wallis test results for the percentage of retweets failed to reject the null (chi-squared = 2.0031, $p = 0.9194$) which suggests that all groups users in this study have similar patterns to the proportions of retweets. A pairwise Wilcoxon signed-rank test on the same set of data also fails to reject the null ($p = 1.0$ for all group pairings).

6.4.1.3 Reweets with links (Equation 6.3)

The percentage of celebrity retweets with links ($RTwL_{user}$) had a median value of 5.654%, and a maximum value of 44.200%, compared to a median of 1.228% and maximum of 17.230% for regular users. Once again, most user groups had at least one user who never had retweeted content with a link, every user in the Entertainment and Higher Education communities did so at least once.

Figure 6.4 shows boxplots for the percentage of retweets with links for the sampled users. The Kruskal-Wallis test results for the percentage of retweets with links suggests differences in the groups (chi-squared = 45.6556, $p = 3.467 \times 10^{-08}$). While the pairwise Wilcoxon signed-rank fails to reject the null hypothesis when comparing

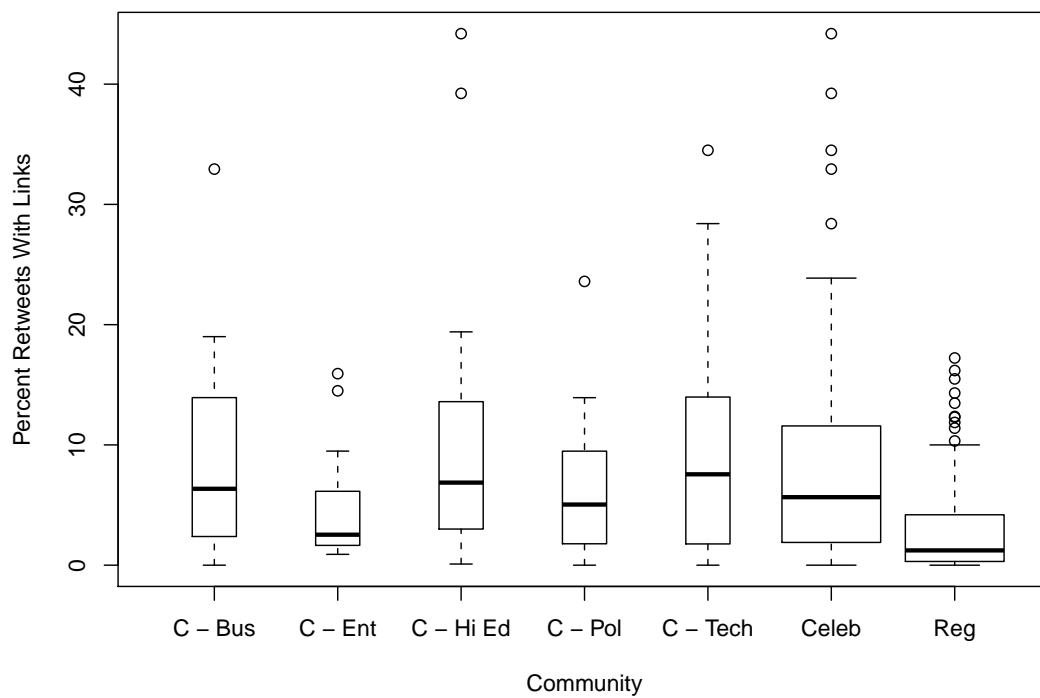


Figure 6.4: Percent Retweets With Links Boxplot

the groups of celebrities ($p > 0.843$), it does show that overall, celebrities are more likely to post retweets with links than regular people ($p = 4.18 \times 10^{-07}$). However, that difference is driven by celebrities from the business, higher education, and technology groups ($p < 0.273$); comparisons between regular users and politicians ($p > 0.084$) or entertainers ($p > 0.892 \times 10^{-3}$) rejected the null.

Results for Hypothesis 1: Rejected

Celebrities and regular users retweet plain tweets at the same rate, rather than less often as supposed by the hypothesis. Additionally, celebrity retweets and plain tweets are more likely to include links to external sites than retweets and plain tweets from regular users. This suggests that celebrities may not be posting original content so much as disseminating content found elsewhere.

6.4.2 Retweet Behavior of the followers

Shapiro-Wilk tests rejected the null hypothesis that the data comes from a normal distribution for each type of retweeted tweet, shown in Table 6.9. Figure 6.5 shows the breakdown of the types of tweets that were harvested broken down by the groups of users. Section 6.4.2.1 provides an overview of the follower's behavior across all tweet types; subsequent sections focus on a particular type of tweet and how often followers retweeted that type.

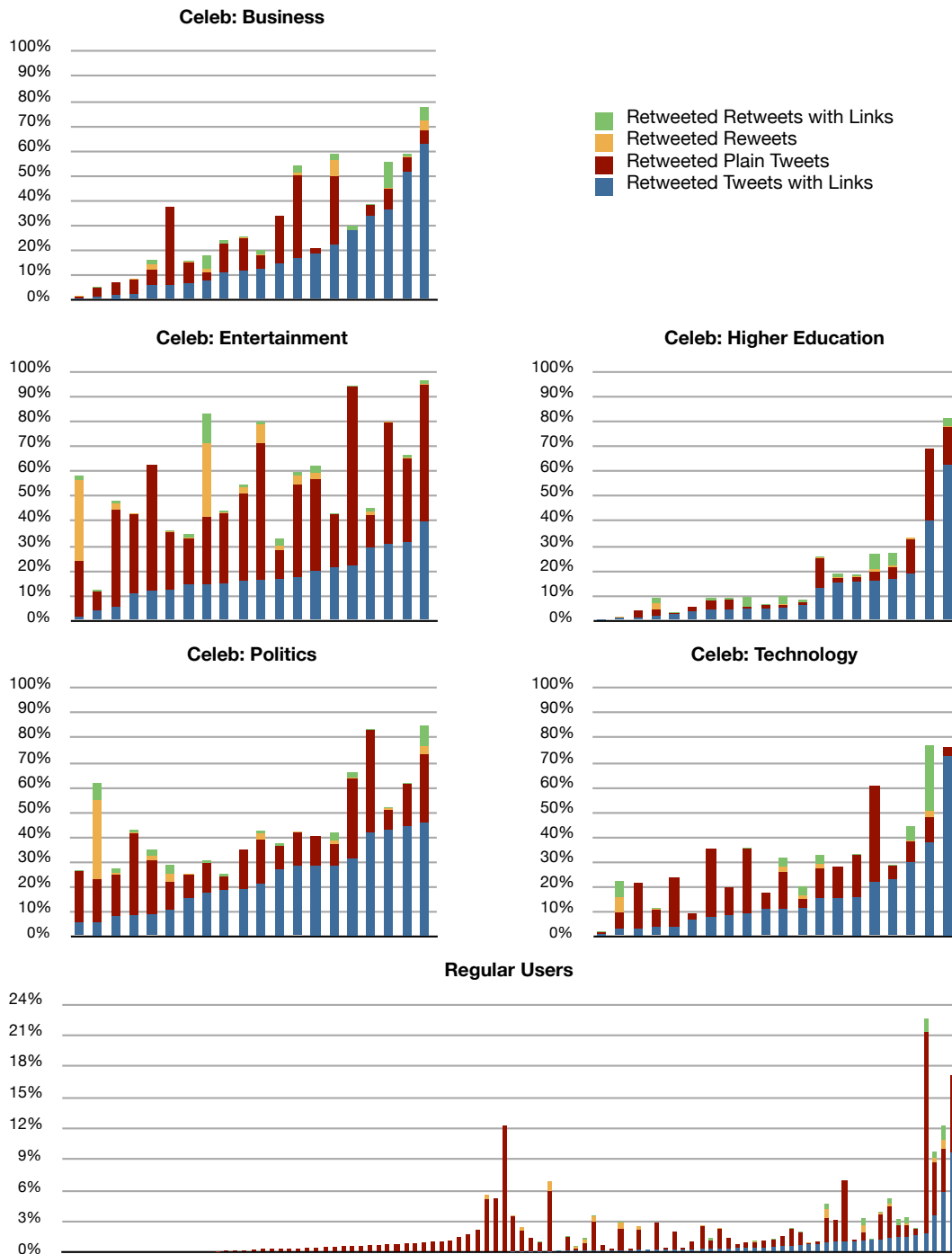


Figure 6.5: Retweeted Tweet Distributions for each category of user. Each user is represented by a bar along the x-axis, sorted by Retweeted Tweets With Links. The empty space above each bar to 100% represents the user's tweets that were not retweeted.

Table 6.9: Shapiro-Wilk Analysis of Retweeted Tweet Types

Tweet Type	W	<i>p</i>-value (<i>i</i>=)
Retweeted Plain Tweets ($RTed\ T_{user}$)	0.8653	1.634×10^{-15}
Retweeted Retweets ($RTed\ RT_{user}$)	0.2542	2.2×10^{-16}
Retweeted Tweets with Links ($RTed\ TwL_{user}$)	0.785	2.346×10^{-12}
Retweeted Retweets with Links ($RTed\ RTwL_{user}$)	0.4563	2.2×10^{-16}
Retweeted Tweets Overall ($RTed\ OT_{user}$)	0.7785	2.2×10^{-16}

6.4.2.1 Retweeted Tweets Overall (Equation 6.8)

The median value for the percentage of celebrity tweets that were retweeted by their followers ($RTed\ OT_{user}$) was 32.639%, and the maximum value was 96.033%, compared to a median of 0.895% and maximum of 22.511% for regular users. At least one regular user was never retweeted, but all celebrities were retweeted at least once.

Figure 6.6 shows boxplots for the percentage of retweets for the sampled users. The previous results, a visual inspection of the boxplots, and the Kruskal-Wallis test results for the percentage of tweets overall that were retweeted (chi-squared = 194.03, $p < 2.2 \times 10^{-16}$) suggests differences in the behavior of followers among the groups of users. Pairwise Wilcoxon tests show that overall, regular users' tweets are far less likely to be retweeted than a celebrity's tweet ($p < 3.196 \times 10^{-07}$) whether split into their individual groups or taken as a whole. Entertainers are more likely to be retweeted than any group of celebrities ($p < 8.808 \times 10^{-03}$) except politicians ($p = 0.594$). Likewise, politicians are more likely to be retweeted than those in higher

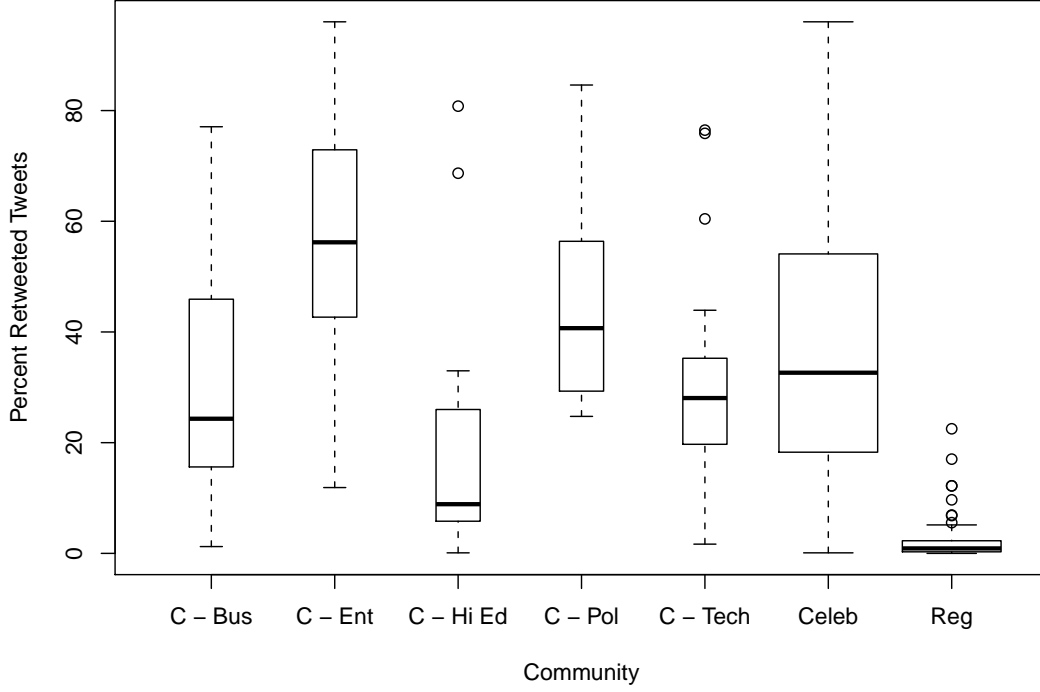


Figure 6.6: Percent Retweeted Tweets Overall Boxplot

education ($p = 3.286 \times 10^{-04}$). For all other pairwise combinations of celebrity groups, the null was rejected ($p > 0.107$), indicating little or no difference in the frequency at which they are retweeted by their followers.

6.4.2.2 Retweeted Plain Tweets (Equation 6.4)

The percentage of celebrity plain tweets that were retweeted by their followers ($RTed\ T_{user}$) had a median value of 11.667%, and a maximum value of 71.616%, compared to a median of 0.627% and maximum of 19.481% for regular users. At least one

celebrity from each of the Business and Higher Education communities never posted a plain tweet that was subsequently retweeted by their followers.

Figure 6.7 shows boxplots for the percentage of retweeted plain tweets (where the original tweet is not a retweet and does not contain a link) for the sampled user's followers. The Kruskal-Wallis test results for the percentage of plain tweets that were retweeted by followers (chi-squared = 167.0398, $p < 2.2 \times 10^{-16}$), along with visual inspection of the boxplot, suggests that all groups of users in this study have similar patterns to the proportions of retweets. The pairwise Wilcoxon signed-rank tests reveal that regular users are the least likely to have their plain tweets retweeted by their followers ($p < 0.0108$) and that entertainers are the more likely to have their plain tweets retweeted than any other group ($p < 1.57 \times 10^{-2}$). It also shows no significant difference between technology and politics or technology and business ($p = 1.0$). Likewise, business users' plain tweets are retweeted at similar rates as politicians ($p = 0.0503$) and higher education users ($p = 0.214$). Higher education users' plain tweets are less likely to be retweeted than politicians ($p = 3.79 \times 10^{-4}$) or technologists ($p = 0.0258$).

6.4.2.3 Retweeted Tweets With Links (Equation 6.6)

The median value for the percentage of celebrity tweets with links that were retweeted by their followers ($RTedTwL_{user}$) was 14.255%, and the maximum value was 72.370%, compared to a median of 0.017% and maximum of 9.610% for regular users. All celebrities posted at least one tweet with a link that was retweeted.

Figure 6.8 shows boxplots for the percentage of retweeted tweets with links for

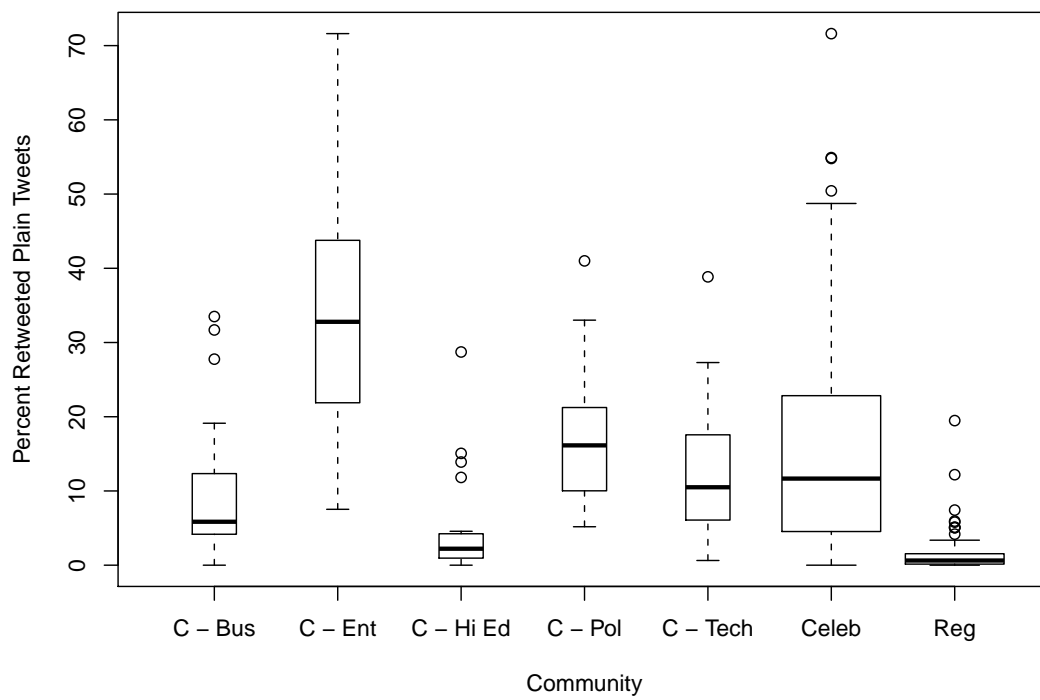


Figure 6.7: Percent Retweeted Plain Tweets Boxplot

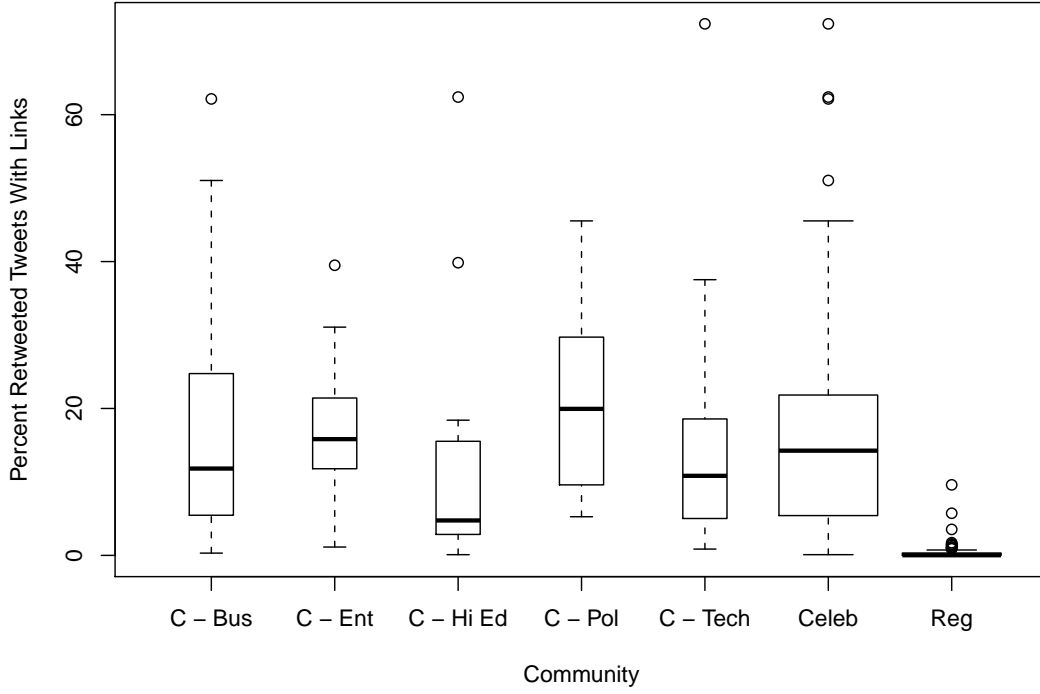


Figure 6.8: Percent Retweeted Tweets With Links Boxplot

the sampled user's followers. Again, visual inspection suggests differences between the groups, and the Kruskal-Wallis test results for the percentage of retweeted tweets with links shows the same (chi-squared = 191.2746, $p < 2.2 \times 10^{-16}$). Pairwise Wilcoxon tests show that regular users are again retweeted less often than any one group of celebrities or the combination of all of them ($p < 7.351 \times 10^{-10}$), while among celebrities, politicians tweets with links are more likely to be retweeted than those of users in higher education ($p = 2.342 \times 10^{-2}$); there were no other significant differences between celebrities ($p > 0.453$).

6.4.2.4 Retweeted Retweets (Equation 6.5)

The median value for the percentage of celebrity retweets that were retweeted by their followers ($RT_{ed} RT_{user}$) was 0.200%, and the maximum value was 32.667%, compared to a median of 0.000% and maximum of 0.924% for regular users. Only celebrities from the Entertainment community all had at least one of their retweets later retweeted by their followers; all other groups of celebrities had at least one person who never posted a retweet that was subsequently retweeted by their followers.

Figure 6.9 shows boxplots for the percentage of retweeted retweets (posts that the sampled users forwarded to their followers, which were then forwarded again by the followers) for the sampled user's followers. While the outliers in the data make it difficult to visually examine the differences between groups in the boxplots, the Kruskal-Wallis test results for the percentage of retweeted retweets suggests there are differences (chi-squared = 83.0054, $p = 8.544 \times 10^{-16}$). Once again, the pairwise Wilcoxon tests show that regular users are retweeted less often than any group of celebrities, whether split or combined ($p < 0.169 \times 10^{-2}$). However there were no significant differences between the groups of celebrities ($p > 0.0756$).

6.4.2.5 Retweeted Retweets With Links (Equation 6.7)

The median value for the percentage of celebrity retweets with links that were retweeted by their followers ($RT_{ed} RT_{wL_{user}}$) was 0.821%, and the maximum value was 26.367%, compared to a median of 0.000% and maximum of 1.362% for regular users. Only celebrities from the Entertainment community all had at least one of their retweets

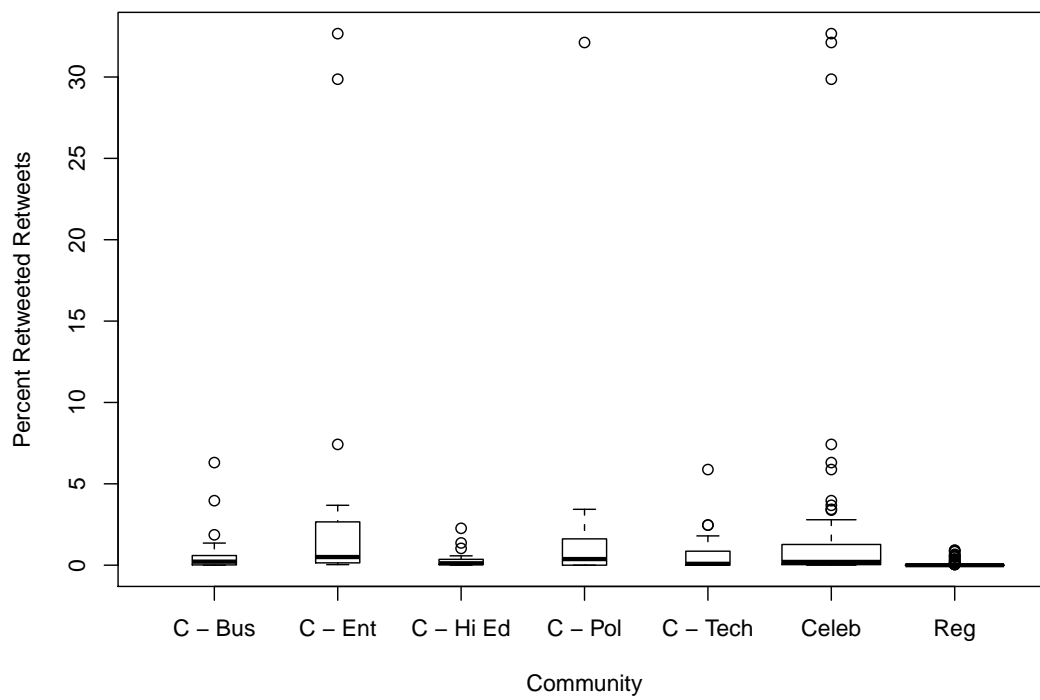


Figure 6.9: Percent Retweeted Retweets Boxplot

with a link later retweeted by their followers; all other groups of celebrities had at least one person who never posted a retweet with a link that was subsequently retweeted by their followers.

Figure 6.10 shows boxplots for the percentage of retweeted retweets with a link (posts containing links that the users read and forwarded to their followers who subsequently forwarded the links again) for the sampled users. The Kruskal-Wallis test results for the percentage of retweeted retweets with links, along with visual inspection of the boxplots, suggests there are differences in between the comparison groups (chi-squared = 123.4268, $p < 2.2 \times 10^{-16}$) and the pairwise Wilcoxon signed-rank test reveals that once again, regular users are retweeted less often than the celebrities ($p < 2.526 \times 10^{-5}$) for individual celebrity groups and taken as a whole.

Results for Hypothesis 2: Supported

Regardless of tweet type, regular users are less likely to be retweeted than celebrities. For most types of tweets, celebrities from the Entertainment community are more likely to be retweeted than those from other communities.

6.4.3 Tweet Content

Of the 3,000 tweets classified, the examiners agreed on the overall category for 1,884 tweets (62.8% agreement). The counts of each category for the different users groups are plotted in Figure 6.11. Pearson's Chi-Squared test on the first view, com-

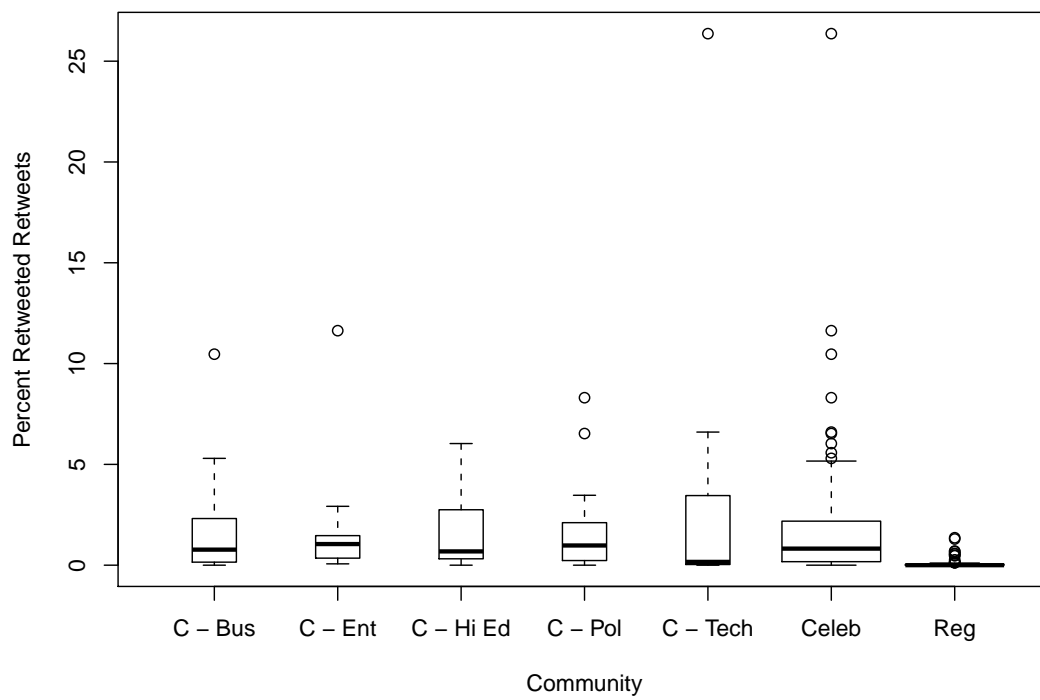


Figure 6.10: Percent Retweeted Retweets With Links Boxplot

paring celebrity users to regular users, rejected the null ($p = 2.11 \times 10^{-9}$); X-squared was 43.3149 with three degrees of freedom. The second view, comparing the different celebrity communities, also rejected the null ($p < 2.2 \times 10^{-16}$); X-squared was 150.2151 with 12 degrees of freedom. In both cases, the observed distribution of tweet categories differed between the comparison groups.

Results for Hypothesis 3: Inconclusive

There is support for the hypothesis that celebrities post more work-related tweets than regular users in that celebrities post more news related tweets and fewer personal tweets than regular users, but the results here are not conclusive. However, the celebrity categorization results are more interesting. Entertainers, are far more likely to tweet about their personal lives than other celebrities, and far less likely to post news related tweets. Politicians, on the other hand, post the most news related tweets by far, and offer the least commentary.

6.4.4 Tweet Timing

The pairwise Kolmogorov-Smirnov tests on the time of day data for all users shows different rates of tweets among most of the tweet types ($p < 1.179 \times 10^{-05}$), but shows that links and tweets that were retweeted have similar distributions ($p = 0.8928$), as do plain retweets and retweets with links ($p = 0.4413$); these results matched those found when looking only at celebrity tweets. For regular users, the distributions are

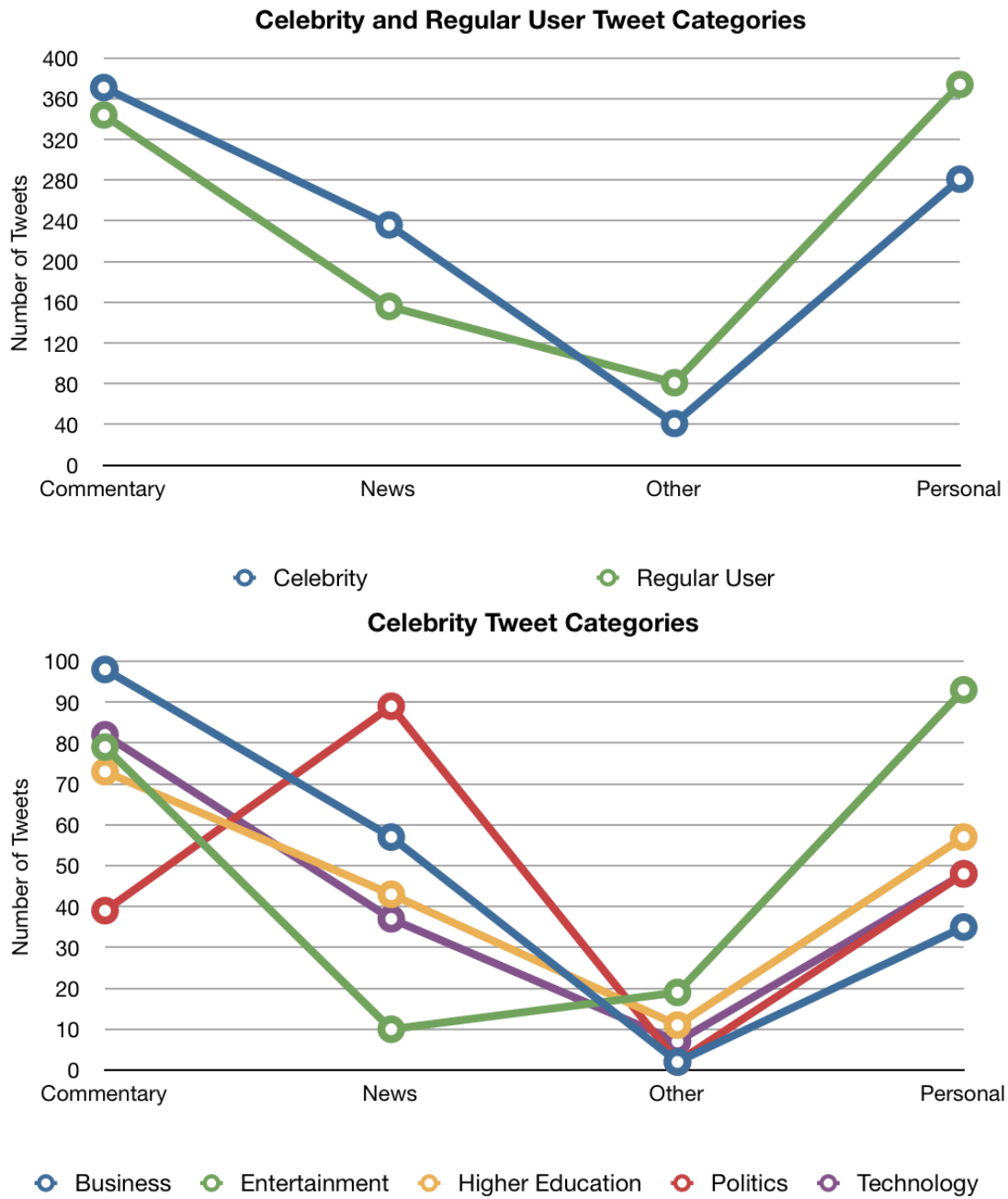


Figure 6.11: Tweet Content Categories

different, with $p < 1.179 \times 10^{-05}$ for all pairs except plain retweets and tweets with links which had a much higher, but still had a significant result of $p = 0.004958$.

Figures 6.12 through 6.14 show the number of tweets posted during a particular hour of the day. Note that the tweet's creation time is reported at GMT, so depending on the time of year, midnight in New York City is at 5 o'clock or 6 o'clock in GMT, while midnight in San Francisco is at 8 o'clock or 9 o'clock in GMT. Assuming that most people get the majority of their sleep between midnight and 6 o'clock in the morning, there should be a significant drop in the number of tweets posted by people in the United States during those hours, which is reflected in these plots.

The pairwise Kolmogorov-Smirnov tests on the day of week data for all users shows different rates of tweets among most of the tweet types ($p < 0.001824$), but shows that links and tweets that were retweeted have similar distributions ($p = 1$), as do plain retweets and retweets with links ($p = 0.9375$); these results matched those found when looking only at celebrity tweets. For regular users, most of the distributions are different ($p < 0.001824$), except retweets with links and tweets that were retweeted ($p = 0.01168$).

Figures 6.15 through 6.17 show the number of tweets posted on a particular day of the week. Again, these days are based on times in GMT, but given the majority of tweets from users in the United States are not posted in the early morning, this should have little impact on the plots here. There is a drop in tweets on the weekends, which could be a result of people spending less time in front of a computer than they do during the business week.

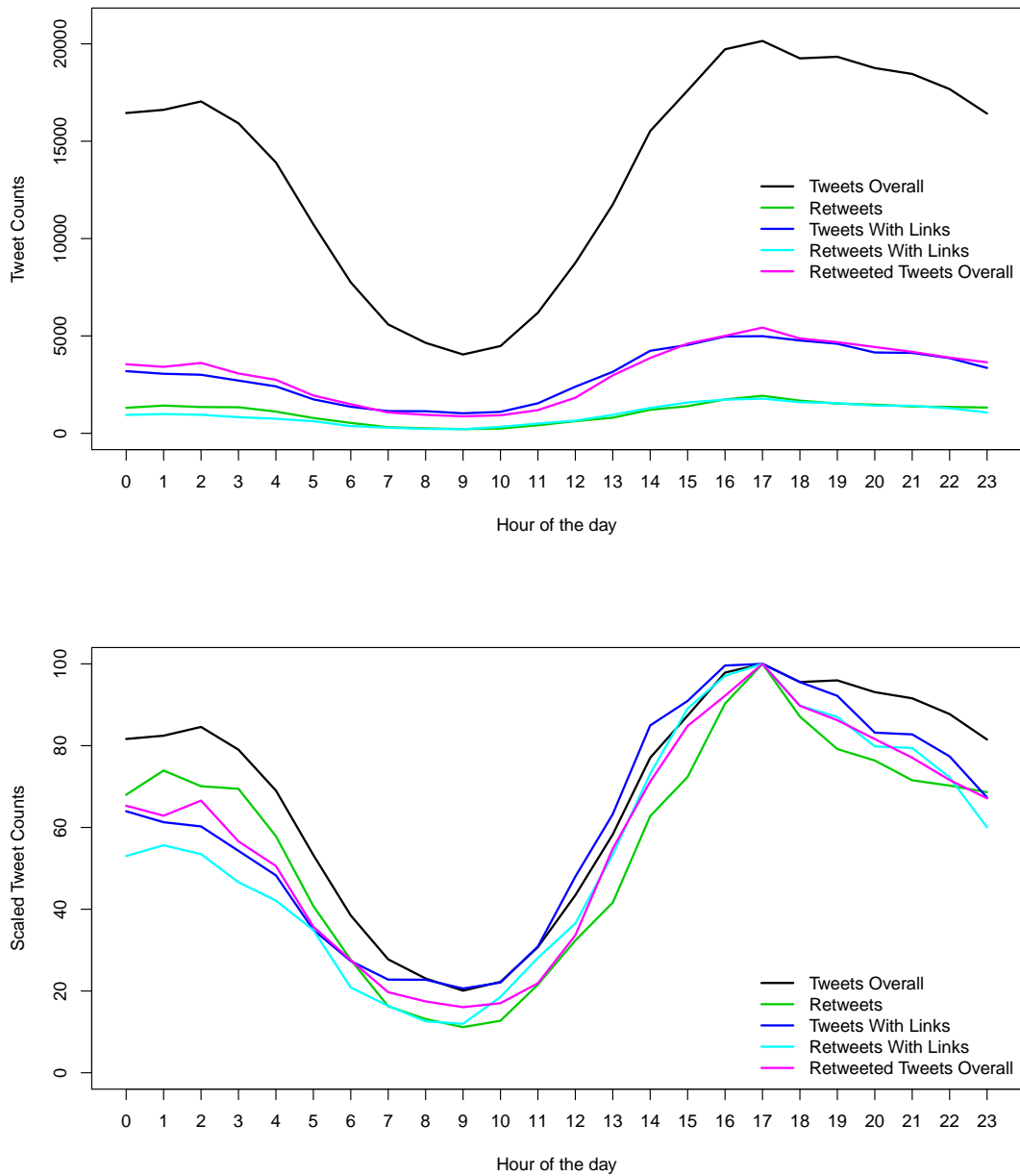


Figure 6.12: Hour of Day (GMT): All Users – The number of tweets posted during a particular hour of the day (GMT) for celebrities and regular users. The data points are non-continuous.

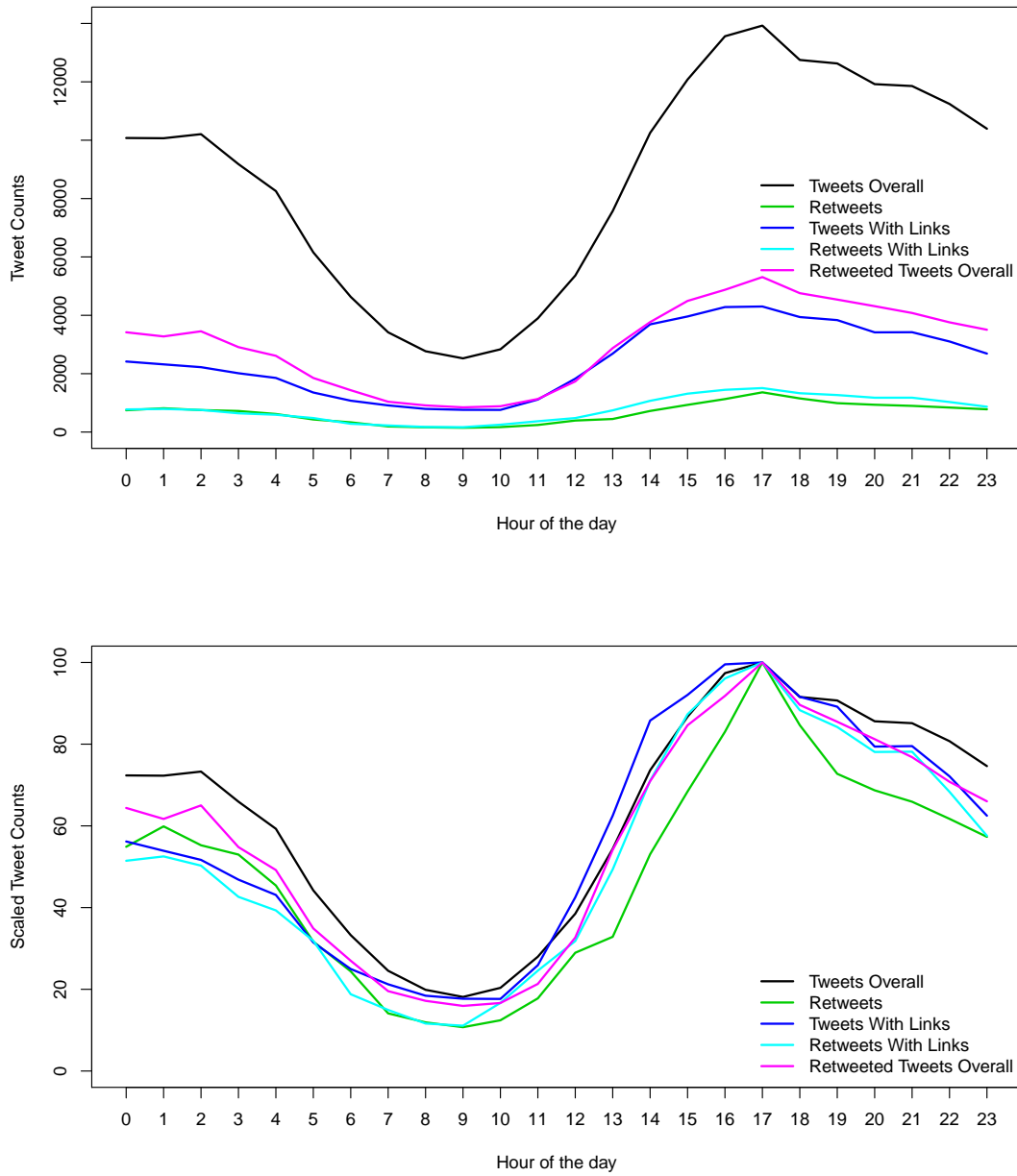


Figure 6.13: Hour of Day (GMT): Celebrities – The number of tweets posted during a particular hour of the day (GMT) for celebrities. The data points are non-continuous.

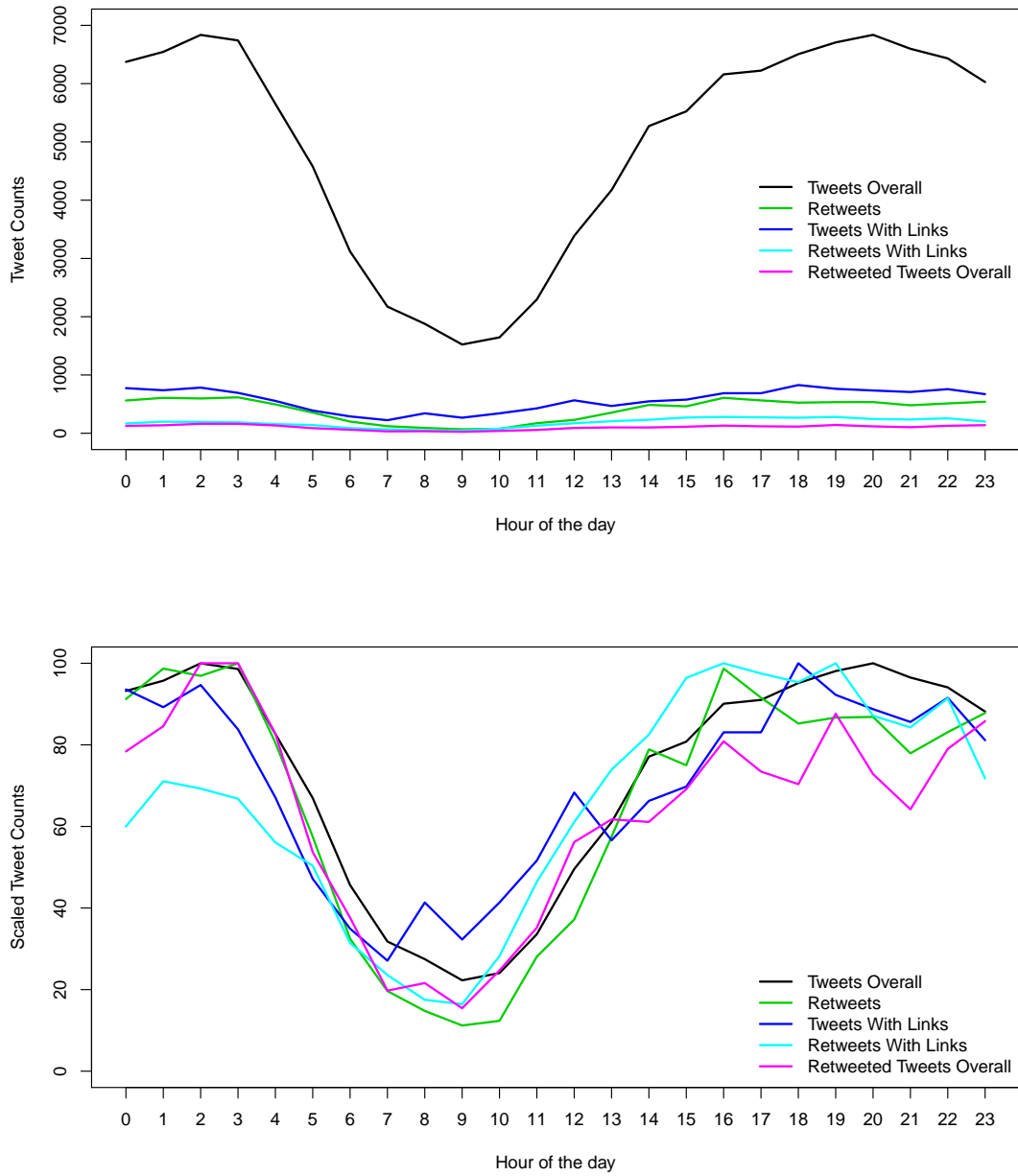


Figure 6.14: Hour of Day (GMT): Regular Users – The number of tweets posted during a particular hour of the day (GMT) for regular users. The data points are non-continuous.

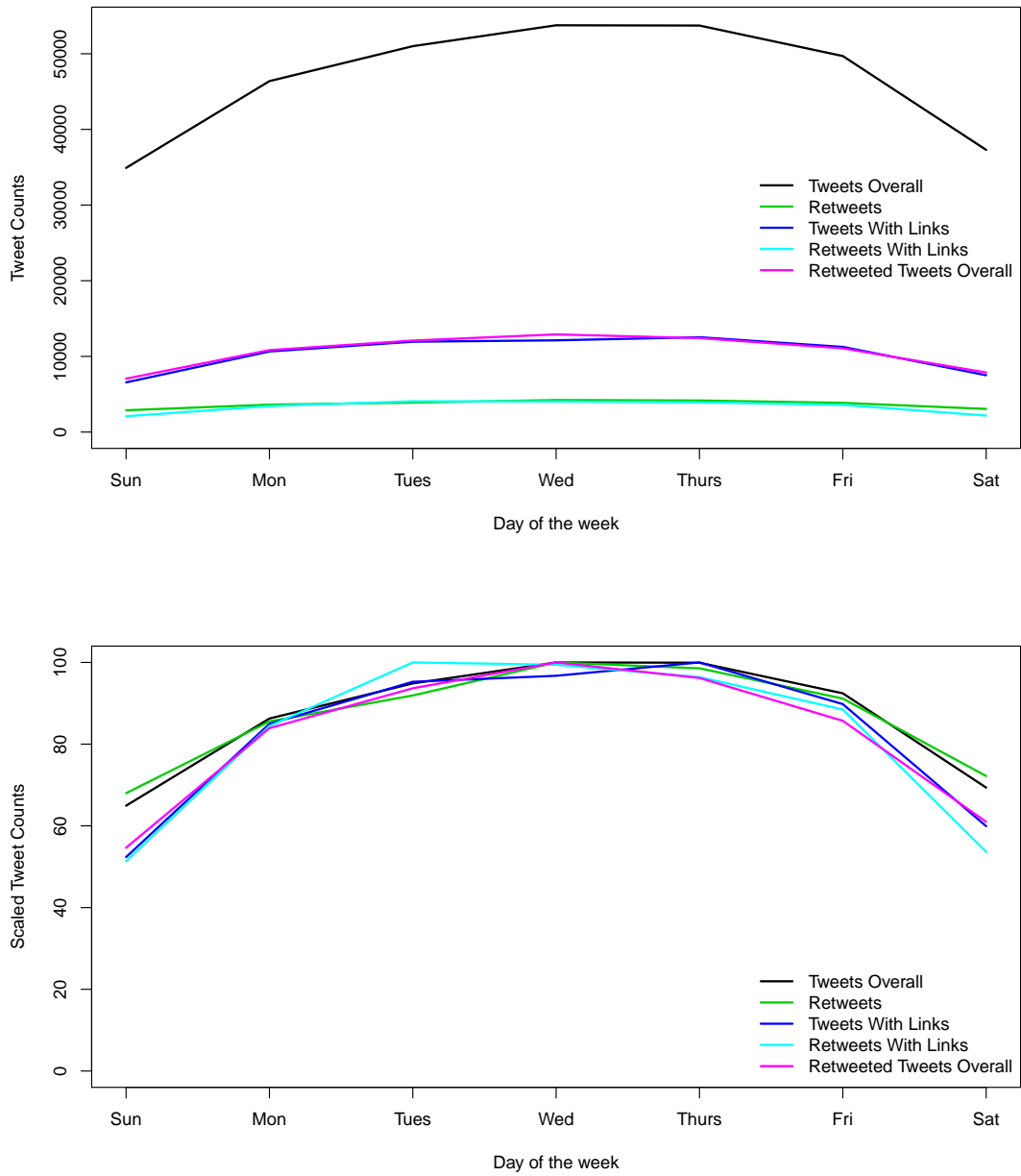


Figure 6.15: Day of Week: All Users – The number of tweets posted on a particular day of the week (GMT) for celebrities and regular users. The data points are non-continuous.

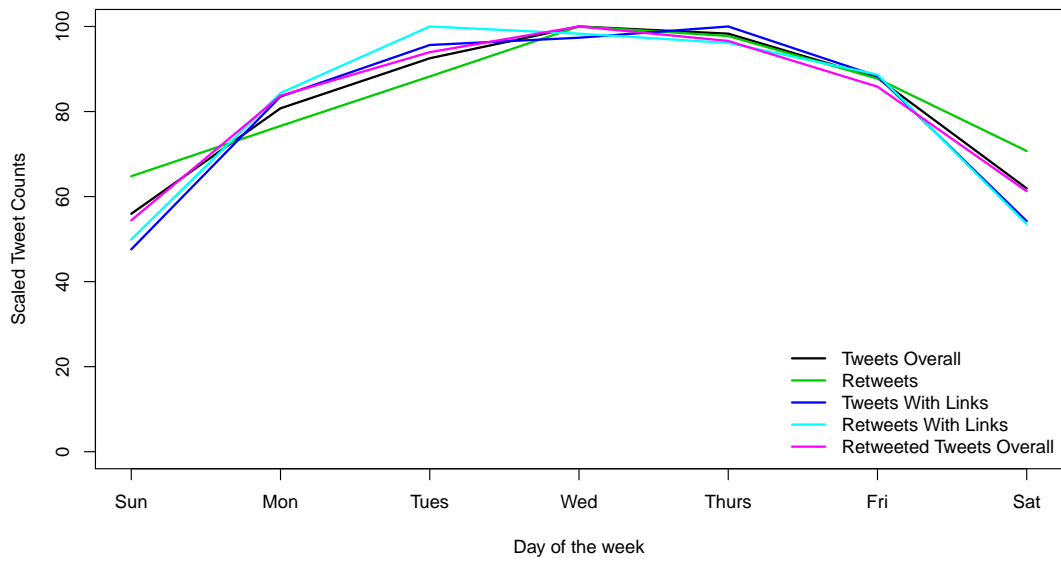
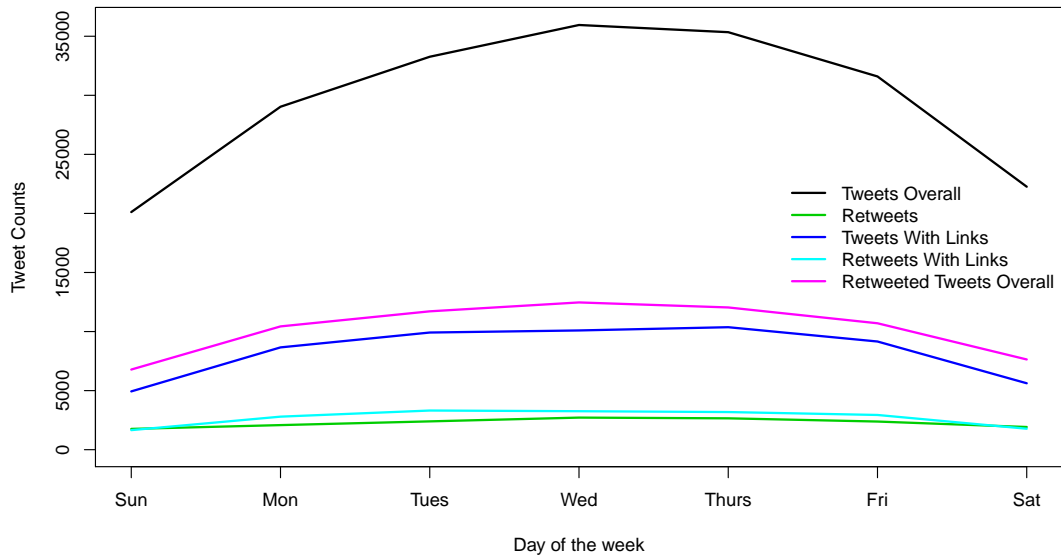


Figure 6.16: Day of Week: Celebrities – The number of tweets posted on a particular day of the week (GMT) for celebrities only. The data points are non-continuous.

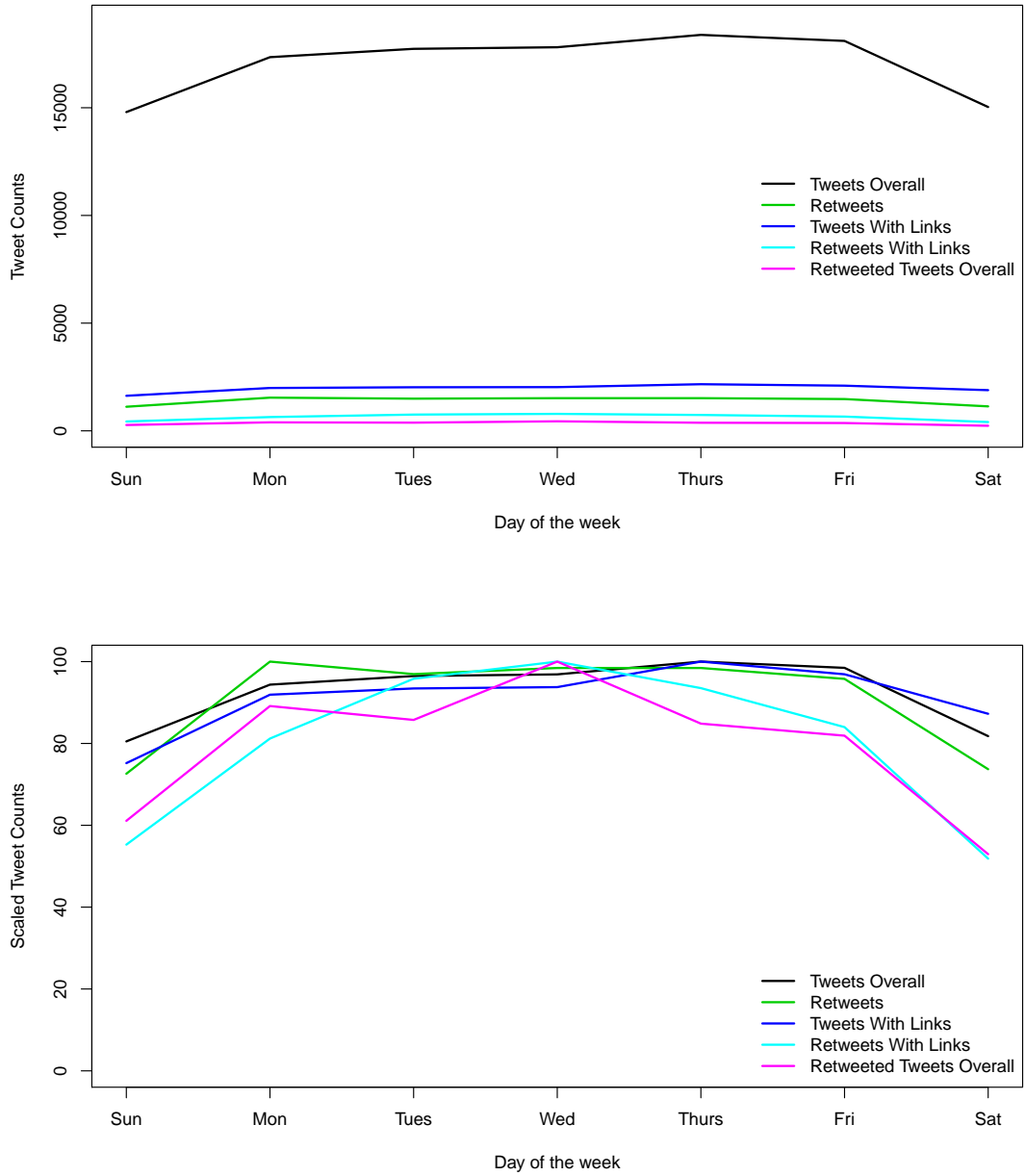


Figure 6.17: Day of Week: Regular Users – The number of tweets posted on a particular day of the week (GMT) for regular users only. The data points are non-continuous.

Results for Hypothesis 4: Supported

The data clearly shows that, for the users in this study, the rate at which any type of tweet is posted in a given hour follows the same pattern, with nearly identical peaks and troughs. The pattern follows the standard work and sleep schedule in North America, where most of the users in the study live, with tweets dropping off between 12 AM EST and 6 AM EST, picking up in the morning and growing steadily throughout the work day, and coming to a peak in the early evening around 6 PM EST, and then again in the late evening around 10 PM EST. The dual peak seen here could be indicative of the different timezones on the east and west coast; since the majority of the North American population lives on either coast, there may be significantly more Twitter users in the EST and PST timezones than in the CST and MST timezones, causing a slight drop in the overall tweet rate between the end of the work day on the two coasts.

Results for Hypothesis 5: Supported

The data clearly shows that, for the users in this study, the rate at which any type of tweet is posted on a given day follows the same pattern, with nearly identical peaks and troughs. Users are far more likely to be active on Twitter during the standard Monday through Friday work week, than on the weekends.

6.4.5 Tweet Rate and Followers

The box plot in Figure 6.18 shows the distribution of tweet rates across all users in the study, both celebrities and regular users. The mean tweet rate was 19.023 tweets per day, and the median was 1.691 tweets per day. There is a positive correlation between the percentage of retweets from followers and the user's tweet rate for celebrity users ($Z = 5.51$, degrees of freedom = 98, $p < 0.001$, see the red fit line in Figure 6.19) and over all users ($Z = 10.01$, degrees of freedom = 198, $p < 0.001$, see the black fit line in Figure 6.19). However, regular users did not show a significant effect ($Z = 1.52$, degrees of freedom = 98, $p = 0.13$, see the blue fit line in Figure 6.19). The effect across all users is likely driven by the very strong effect in celebrities. While there is an increase in the percentage of tweets that a celebrity's followers will retweet, there is no such response for regular users.

Results for Hypothesis 6: Partially supported

Only celebrities will see an increased percentage of retweets by posting tweets more often; regular users will not.

The box plot in Figure 6.21 shows the distribution of the number of followers across all users in the study, both celebrities and regular users. The mean number of followers was 219544, while the median was 1300 followers; celebrities, obviously, have many more followers than regular users. There is a positive correlation between the percentage of retweets from followers and the number of the user's followers over

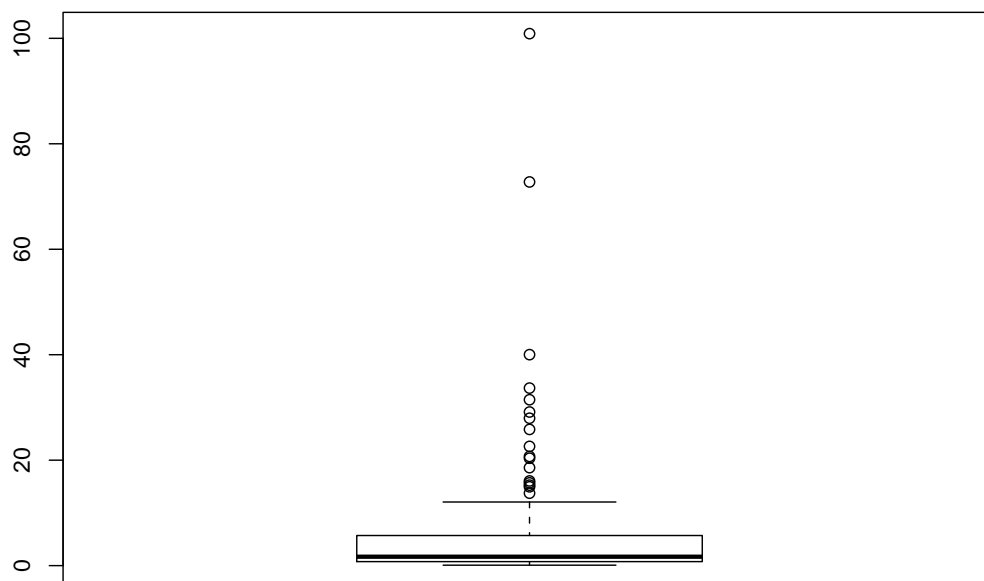


Figure 6.18: Tweet Rate: Tweets per Day

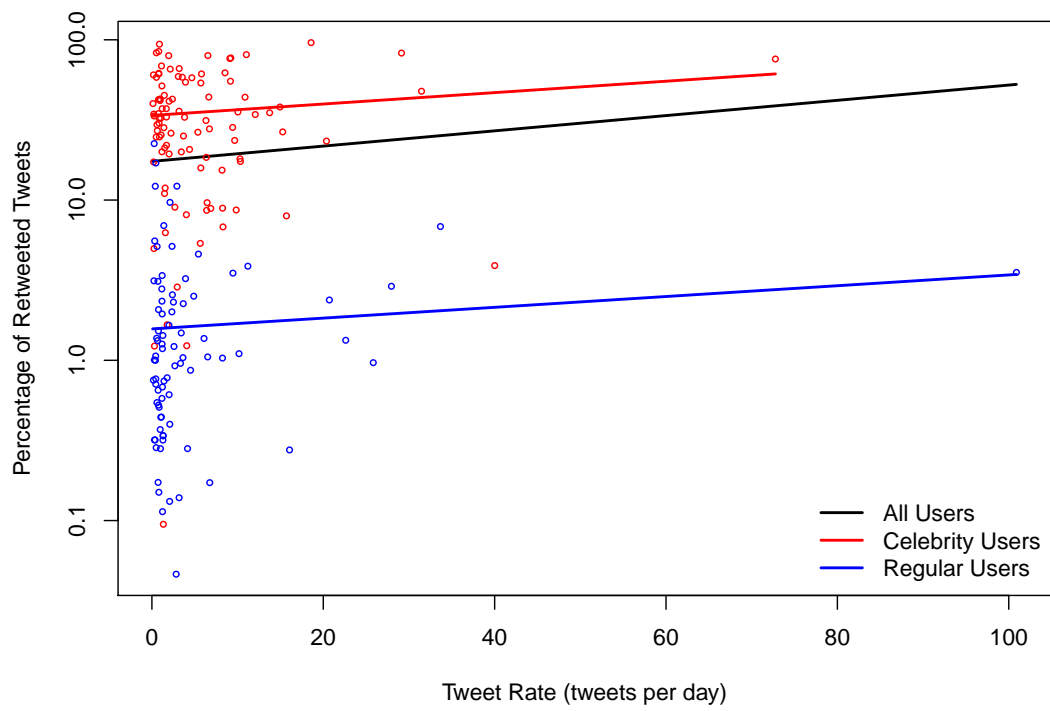


Figure 6.19: Tweet Rate Regression

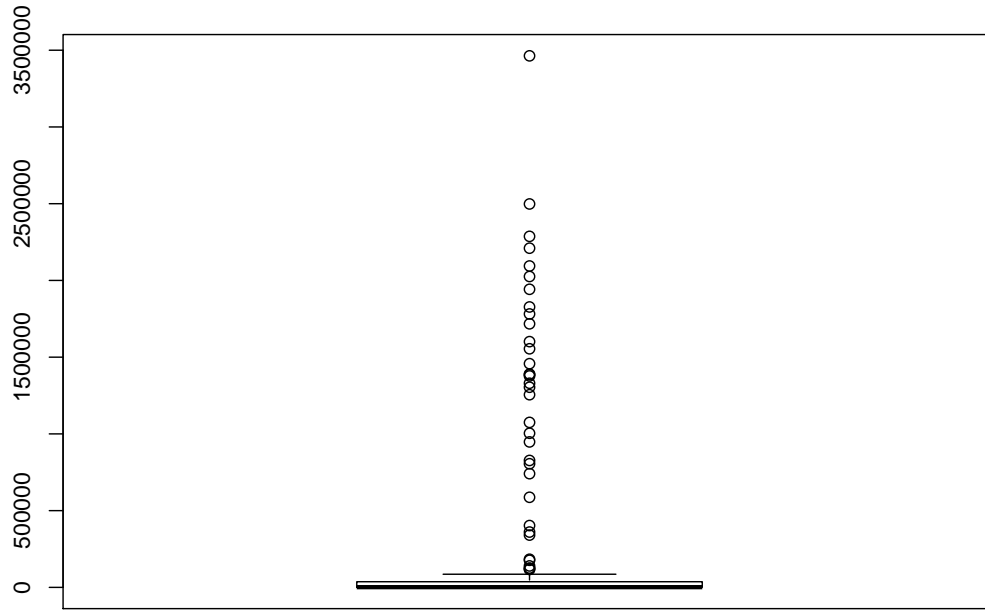


Figure 6.20: Follower Count

all users ($Z = 41.75$, degrees of freedom = 198, $p < 0.001$, see the black fit line in Figure 6.21). Unlike a user's tweet rate, the effect here is also positive for both celebrity users ($Z = 17.97$, degrees of freedom = 98, $p < 0.001$, see the red fit line in Figure 6.21) and regular users ($Z = 14.78$, degrees of freedom = 98, $p < 0.001$, see the blue fit line in Figure 6.22).

Results for Hypothesis 7: Supported

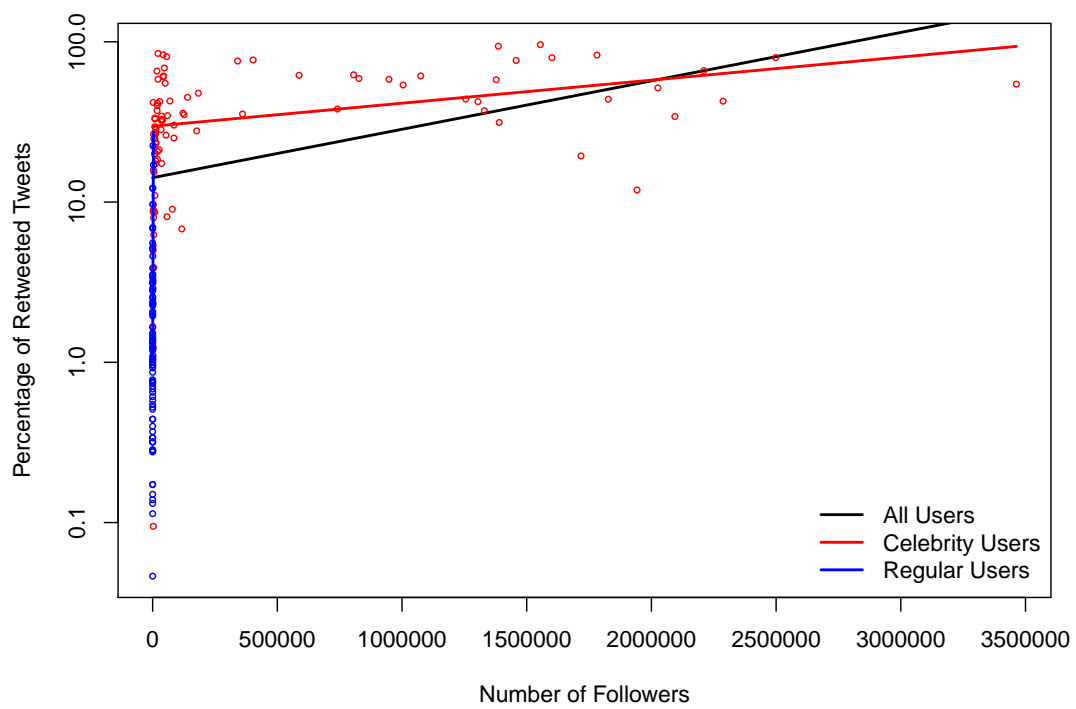


Figure 6.21: Number of Followers (all users)

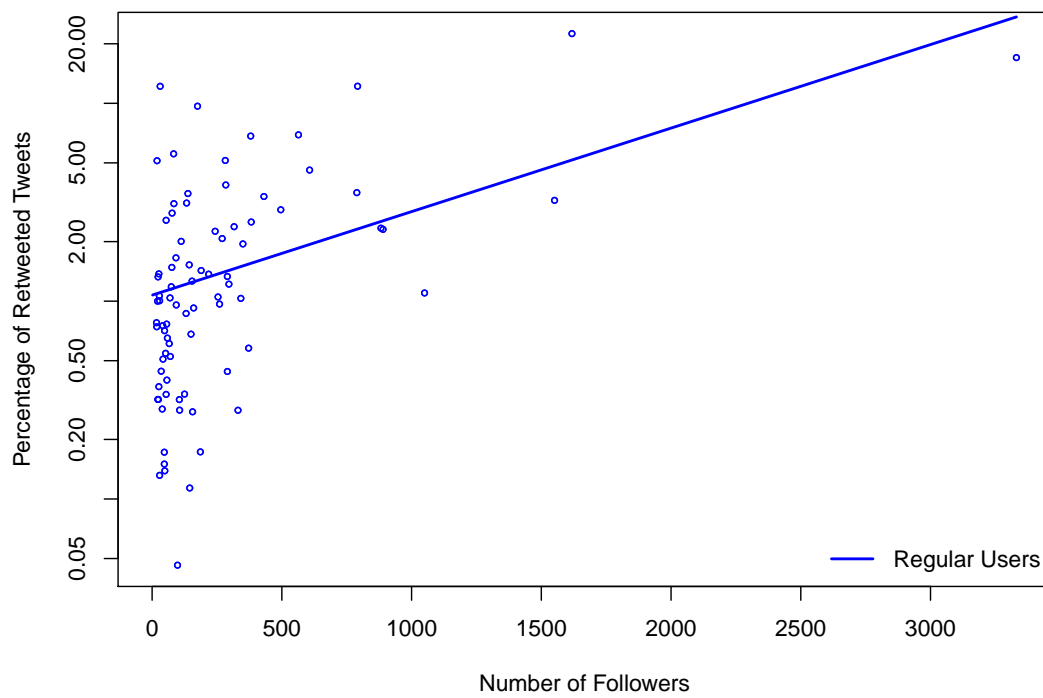


Figure 6.22: Number of Followers (regular users only)

Both celebrities and regular users will see an increase in the percentage of their tweets that are retweeted by accumulating more followers.

6.5 Discussion

This section discusses the results and what was learned in this study. Section 6.5.1 explores the findings related to knowledge transfer, Section 6.5.2 looks at findings related to tweet content, and Section 6.5.3 examines patterns seen in the times that tweets are posted. Finally, Section 6.5.4 discusses threats to the validity of these results.

6.5.1 Knowledge Transfer via Tweets

The first result (reported in Section 6.4.1) is also the most surprising — all groups of users examined in this study retweet at about the same rate, refuting Hypothesis 1. Additionally, when retweets are broken down by those with links and those without, celebrities were found to retweet more tweets with links than the regular users. Less surprisingly, celebrities are also more likely to include links in their tweets than regular users. Taken together, these findings suggest that celebrities are a better conduit than regular users for bringing information from the greater web and disseminating it on Twitter.

Among the least surprising findings is that regular users are retweeted less than celebrities, and when regular users are retweeted, their plain tweets are retweeted

more often than any other type. In contrast, celebrity tweets are far more likely to be retweeted if they contain a link. And while one might expect something that has been retweeted once to be more interesting, or have higher quality information, thereby making it more likely to be retweeted again, the results here found just the opposite — retweets of any kind are very unlikely to be retweeted again; this effect mirrors the results Wu et al. [170] found in their study of information flow via email messages.

Predicting the spread of a particular tweet depends on knowing the number of followers the user has (F), as well as the expected rate of retweets for that user and type of tweet (R_e). We refer to the product of these two values (Equation 6.11) as the Coefficient of Retweetability (C_R). Based on the results reported in Section 6.3.4.3, we can calculate Coefficients of Retweetability for a fictitious celebrity with 50,000 followers and a fictitious regular user with 100 followers. The follower numbers given to the model users are slightly higher than the median number of followers for the users in this study (36,808 followers for the celebrities and 85 followers for the regular users); this has been done to make the calculations easier to follow.

$$C_R = F \times R_e \quad (6.11)$$

Table 6.10 shows the calculated C_R values for the model celebrity and regular user for each type of tweet examined in this study; the median retweet rate is the same reported in Section 6.3.4.3 for the respective type of user and type of tweet. Note that the C_R for the regular user is zero for anything other than plain tweets. This allows us to predict the expected distribution for a generic tweet from our model users, because we

Tweet Type	Median Retweet Rate		Calculated Coefficient of Retweetability	
	Celebrity Users	Regular Users	Celebrity (50,000 Followers)	Regular User C_R (100 Followers)
Plain Tweet	11.667%	0.627%	5834	1
Tweet with Link	14.255%	0.017%	7128	0
Retweet	0.200%	0.000%	100	0
Retweet with Link	0.821%	0.000%	411	0

Table 6.10: Computed Coefficients of Retweetability

	Plain Tweet	Retweet
Distribution =	50,000	$+ (CelebC_R \times 100) + (RegC_R \times 100)$
Distribution =	50,000	$+ (5,834 \times 100) + (0 \times 100) = 633,400$

Table 6.11: Expected celebrity tweet audience

	Plain Tweet	Retweet
Distribution =	100	$+ (RegC_R \times 100) + (RegC_R \times 100)$
Distribution =	100	$+ (1 \times 100) + (0 \times 100) = 200$

Table 6.12: Expected regular user tweet audience

can be reasonably sure that unless the tweet is seen and retweeted by another celebrity, there will be at most two steps between the original tweeter and the leaf users who don't retweet it. For the purposes of this demonstration, we assume that none of the model's celebrity users have followers that are also celebrities.

Table 6.11 shows the process for calculating the distribution of a plain tweet from the model celebrity. The tweet is visible to all of the celebrity's 50,000 followers, who are regular users. Because the tweet is from a celebrity, the Celebrity C_R for a plain tweet is used as a multiplier for the regular user's 100 followers, resulting in 583,400 additional users who receive it as a retweet. In the final clause, the retweet is from a regular user so the Regular User C_R is used as a multiplier for the 100 followers, but since that multiplier is 0, there are no additional users who receive the celebrity's tweet

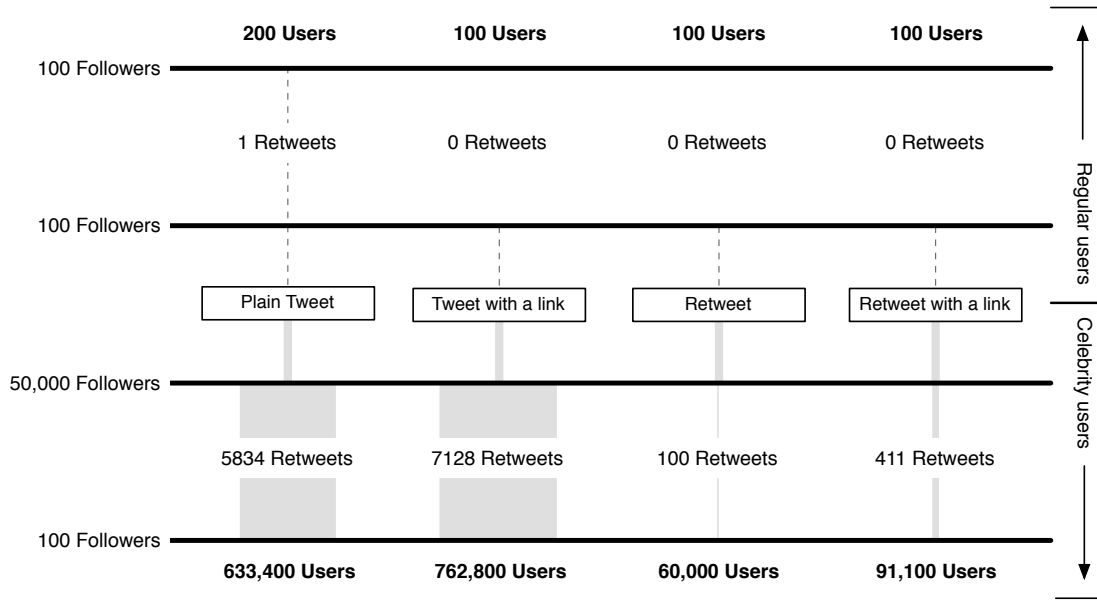


Figure 6.23: A model of knowledge transfer via tweets

as a retweeted retweet. This gives a final audience of 633,400 users for the celebrity’s plain tweet.

Likewise, table 6.12 shows the process for calculating the distribution of a plain tweet from the model regular user. The tweet is visible to all of the user’s 100 followers, who are also regular users. Because the tweet is from a regular user, the Regular User C_R for a plain tweet is used as a multiplier for the regular user’s 100 followers, resulting in 100 additional users who receive it as a retweet. And again, the final clause yields no additional users who receive a retweeted retweet. This gives a final audience of 200 users for the regular user’s plain tweet.

Figure 6.23 presents a model of knowledge transfer via tweets based on the

results reported in Sections 6.4.1 and 6.4.2. The calculations are the same as the one demonstrated in Table 6.11. The different types of tweets captured in this study are in the middle, with two levels of indirection of transfer above and two levels of indirection below. The levels of indirection above the tweet types are for tweets posted by regular users, and the levels of indirection below the tweet types are for tweets posted by celebrities. Each level of indirection is represented by a horizontal line, with the number of Twitter users who might see the tweet at that level on the left. The bars between emanating vertically from the types of tweets represent the size of the readership at each level of indirection. For celebrity tweets, the width of the bars is proportional to the number of potential readers at that level of indirection. For regular users, bars are dashed because the number of readers is so small that the lines would not be visible if drawn at the appropriately scaled width. Only two levels of indirection are displayed because the data gathered in this study indicates that, generally speaking, regular users do not retweet something that was already retweeted by another regular user; once the transfer passes beyond a celebrity's direct followers, it stops spreading very quickly.

At the top and bottom of each tweet type's knowledge transfer flow is a count of the number of potential users who might have been reached by that type of tweet. The model clearly shows that regular users are unlikely to be able to effectively spread knowledge beyond their immediate followers. On the other hand, any tweet posted by a celebrity is likely to be seen by more Twitter users than follow that celebrity directly, even for retweets and retweets with links, which each have well under a 1% chance of being retweeted again.

The model does not directly include the effects of a celebrity user retweeting another celebrity user; celebrities make up a tiny portion of the twitter population so the model represents the most common case. However, for every celebrity tweet that is retweeted by another celebrity, the model indicates that the number of people that tweet reaches will grow by another 60,000 to 91,100 users. The model is also unable to account for unattributed knowledge transfer, such as might occur when a user sees a link posted by someone they follow, and posts their own tweet with that link without retweeting or mentioning who sent the link to them.

Based on the analysis of the effect of tweet rate and increased followers on a user's retweet percentages (reported in Section 6.4.5) and the model presented here, it seems fair to say that Twitter users who want to increase their influence (or at least the percentage of their tweets that are retweeted) should focus on increasing the number of people who follow them. Overcoming the extremely low rate of retweeted retweets, exposing users beyond one level of indirection to a post, seems impossible otherwise.

6.5.2 Tweet Content

The tweet content analysis shows that regular user's retweets, tweets with links, and retweets with links are related to their personal lives more often than celebrities overall, while celebrity's retweets, tweets with links, and retweets with links are more likely to be related to news. Among celebrities, politicians are far more likely to post news items and far less likely to directly engage with their followers by posting tweets containing commentary of some type. Entertainers, however, were far less likely to post

news stories and far more likely to post personal information about themselves.

6.5.3 Tweet Timing

While the distributions of the different types of tweets over the course of the day, and over the course of a week, are generally different, the scaled plots clearly show they follow the same pattern. This pattern is likely due to the relatively low probability that a user will click through more than a few pages of tweet history [88]. The pattern also supports Lerman and Ghosh’s [115] finding that most retweets occur shortly after a tweet is posted; if retweets were more spread out over time, the pattern of retweets would be offset from the pattern of tweets, or would be significantly flatter after being scaled, or both.

However, this finding appears to be at odds with the pattern reported by Zarella [177]; while he noted the same shaped curve reported here for retweets, he did not find that same pattern across all his collected tweets. The users examined here are almost entirely based in the United States, which clearly impacts the pattern seen in the time of day Figures 6.12 through 6.14. Zarella’s study seems to have two different sets of data, one containing “over 40 million ReTweets” and another with a “random sampling of over 10 million “regular” Tweets that may or may not be ReTweets”, but the description of those datasets makes it impossible to tell if that caused the difference in his findings; it could simply be that American Twitter users retweet far more often than those in other countries. The day of the week patterns reported here more closely coincide with the patterns reported by Zarella.

6.5.4 Threats to Validity

The primary threat to the validity of these results comes from the selection of users, both celebrity and regular, that were included in the study. Because celebrity users were selected in an ad hoc fashion, it is possible that they are not a true representation of celebrity behavior, and that their followers are also not a representative sample of celebrity followers. There are fully abandoned accounts that still are counted as followers even though the account owners will never log in again, just as there are accounts that were created only to read tweets, rather than post them; the selection of regular users with “active” accounts screened both of these account types out of the study, which may lead to over-estimations of the percentage of followers who will retweet.

The content analysis portion of this study had five people each read a subset of the tweets, such that each tweet was categorized by two people; only tweets where both people agreed were included in the study. This results of this may have been improved by asking all five people to read all of the tweets, and including tweets where three of the five agreed on the category. Additionally, a larger selection of tweets might have provided clearer results.

All users selected for this study primarily tweeted in English and most lived in North America. It is possible that users in other parts of the world have vastly different behavior, in terms of their likelihood of retweeting, the types of content they post, and even the time of day or day of the week they are most likely to be using Twitter. For example, countries where the majority of users do not have access to the internet while at work might have a very different usage pattern over the course of the day than the

ones presented here; on the other hand the prevalence of mobile phones around the world could make that a moot point. Finally, there are many users who have chosen to keep their accounts and tweets private, so their behaviors could not be included in this study.

6.6 Implications for an Academic Knowledge Transfer System

Status updates and micro-blogging are a relatively new development that combine aspects of speeches (see Section 2.1.2) and electronic messages (see Section 2.1.3). Academics could (and do) provide links to their newly established research projects, as well as their papers that are accepted for publication. They can also share papers from other authors, along with short comments. By re-publishing micro-blog style status updates, the results here clearly indicate that popular users can bring an exponential increase in visibility to the subject of their post. Additionally, the ease with which information can be shared using this style of Electronic Message suggests it can be a useful tool even within a Project group. Status updates, or micro-blogs, would be a welcome addition to any academic knowledge transfer system.

6.7 Conclusion

With the rise of the internet and the culture of constant connection, researchers are, for perhaps the first time, able to passively track the dissemination of knowledge in

real time, or near real time. Twitter is one of many innovative communication channels to appear in the last decade, but its large public dataset and API make it one of the most interesting channels for researchers. Retweets, Tweets with Links, and Retweets with Links are the key measurable knowledge transference mechanism on Twitter. In all three cases, a user had to read a tweet, or see something on a web page, and decide it was interesting enough to pass on to their followers.

This study has shown that celebrities and regular users have very different patterns of usage and follower behavior. Celebrity users seem to embody the *opinion leader* role, posting more knowledge transferring tweets than regular users. Likewise, regular users appear to embody the *media consumer* role, retweeting much of the celebrities' content while having relatively little of their own content retweeted. As a conduit, however, Twitter seems to be deficient in that most knowledge is only passed on one time.

Future work in this area should combine the automatic celebrity account detection employed by Wu et al. [171] to see if the patterns of behavior identified here are seen in a random sample of celebrity accounts, rather than the haphazard selection employed here. With more data, these analyses might be refined to determine which celebrity accounts are corporations or are managed by a team of people, and which ones are real people. Additionally, user studies and interviews could be conducted to explore why retweets are themselves so rarely retweeted.

Chapter 7

Conclusion

This thesis has developed several important insights into how academics share knowledge. This chapter summarizes the thesis' contributions and then suggests areas for future research. It concludes with some final thoughts.

7.1 Contributions

A theoretical model of information sharing, based on knowledge flows and the artifacts that carry knowledge, was developed that provides the foundation for the goals and requirements of a software system that enables academic information sharing to be more efficient. A prototype implementation of one aspect of the software system demonstrates a solid architecture based on the model and software system goals, and validates aspects of the system design. The exploration of gamification for academic knowledge sharing within Facebook demonstrates important pitfalls that should be avoided. Finally, an analysis of information sharing within the Twitter social network provides

insights that apply to design and implementation choices of an academic knowledge sharing platform.

7.1.1 A theoretical model of information sharing

The model developed in this thesis (see Chapter 2) differs greatly from other knowledge transfer models because it focuses on the paths that knowledge must take within the academic research process, while being agnostic with regard to the form (or format) that knowledge takes. The major aspects of academic knowledge transfer are captured with shared repositories for both individuals and groups, support for a number of peer-reviews, and eventual publication with the overall research community (see Section 2.2). This model recognizes that while the artifacts used to convey knowledge may differ between projects and change over time, both in terms of data format and data content, the overall process by which an individual or group performs research and publishes their findings is consistent across media and academic communities. The major advantage of this modeling approach is that it lends itself to the design of a software system built to support the knowledge transfer process.

7.1.2 Goals and requirements for a knowledge sharing software system

The requirements for an academic knowledge sharing system build on the theoretical model and were presented in Chapter 3. Specific artifact repositories for individuals and groups were identified, including digital libraries and versioned file spaces. The digital libraries (Section 3.1.1) allow users to organize source materials and keep their

notes searchable, accessible, and easily sharable. Versioned file spaces (Section 3.1.2) provide an important tool commonly found in software development — but rarely utilized in academic collaborations — that helps prevent data loss when multiple people collaborating on a shared artifact, as well as guarding against accidental changes. The support for projects (Section 3.1.3) can provide both an organizational structure to an individual’s work as well as an environment for close collaboration with others. Community journals (Section 3.1.5) provide support for the peer review process, whether the review is conducted on research proposals or the results of a research project. All artifacts, whether in a digital library, a versioned file space, or community journal, are able to have threaded discussions associated with them. Also, the different work environments (Sections 3.2.2 - 3.2.4) that academics find them in suggest the need for several software clients that are able to access the same data.

7.1.3 A prototype software system

A prototype system for knowledge sharing, detailed in Chapter 4 provided an implementation of the digital library repository and allowed students to use it. This prototype validated the separation of the client interface from the back end system, which would eventually allow clients to be developed across a range of platforms while reusing the back end. The prototype also demonstrated the usefulness of notes and threaded discussions associated with the knowledge artifacts. While the prototype’s user interface left much to be desired (see Section 4.6.2), the feedback from users enabled the development of a more appropriate interface, described in Section 4.7, that could be

appropriate for multiple client platforms.

7.1.4 Gamification of academic knowledge sharing

As the reigning social network, Facebook’s application platform presents an alluring option for the development of an academic knowledge sharing system, such as the one described in Chapter 5. However, Facebook’s general focus on the *social* aspect of its user’s lives makes it an environment that many find inappropriate for use in academic work. For most people, the gamification methods utilized in this study were insufficient to overcome people’s aversion to using Facebook in their work, or for others, encourage them to being connected to collaborators via Facebook or even using Facebook at all.

7.1.5 Information sharing in Twitter and its implications

An analysis of information shared within the Twitter social network (see Chapter 6) provides insights into how knowledge can spread when a person of influence (a celebrity) within a community promotes it. Celebrities in Twitter were more likely to have their messages promoted to other users than non-celebrities; celebrities who post more often are more likely to have their messages promoted as well (Section 6.5.1). For all users in Twitter, the more people that receive a message, the more likely that message will be forwarded to additional users. Additionally, messages shared during the “working week” was more likely to be forwarded to additional users than those posted in the off hours or over the weekend (Section 6.5.3).

These findings have implications for any knowledge sharing system that allows users to subscribe to messages posted (in a “push” method) by other users. Status messages in Facebook and Google+ explicitly fall into that category, as do more traditional email-based mailing lists. A model developed from the data gathered on Twitter (see Figure 6.23) shows that, while messages from regular users might have a readership of several times the number users subscribed to their messages, celebrities can reach an audience that is an order of magnitude greater than the number of people that have explicitly subscribed to their messages. The model also shows that messages commenting on an artifact external to the system are more likely to be forwarded; the difference between the two rates provides a means of measuring and comparing influence between celebrities.

By including this type of subscription and messaging in an academic knowledge sharing system, users will have an easy means of sharing their thoughts on a knowledge artifact within the system. This sharing can both increase awareness of the artifact and its contents, as well as provide the basis for discussions within a community. Perhaps more importantly, it provides a casual counterpoint to the formalized publication process normally used to share with the overall community.

7.2 Future Work

Oleksik et al. noted [131] “one has to provide beyond simple access to common data and metadata” when developing information sharing tools. Their recent study found that academics have very poor tool support for some of the most common tasks

they perform, such as organizing the files (both data and research output) across software applications, and linking research output back to the raw data. Both of these concerns, and others, would be addressed by a knowledge sharing environment such as the one suggested in Chapter 3. The full development of such a system presents opportunities for additional research in human factors, CSCW, and social psychology. Issues surrounding the implementation of the proposed Whisper system have been explored in the preceding chapters suggesting additional features and alterations from the original proposal in Chapter 3. Once a system like Whisper has a large enough user base, it will provide a great deal of additional data for Scientometric-based analysis. Additionally, Scientometric analysis might be applied to the Twitter social network and its tweets and retweets.

7.2.1 Full Scale Implementation

The implementation of the Whisper system (or one based on the same feature set and goals) is a key precursor for much of the future work from this thesis. Such a system, in addition to providing tools to researchers and students, would give researchers an incredible amount of data related to how information flows between people, projects, and research communities. This thesis has explored some of the design paths available for such a system, and has pointed out some pitfalls to be avoided. While others certainly exist, going forward, a prototype of the system with a small set of users will provide higher quality feedback about the design of the system and required feature set than additional small scale exercises such as were undertaken here. This initial development

would be a significant first step in a cycle of continuous improvement as we collectively iterate upon the model and enhance the tools we build to support the processes in which we all operate. At some point, one simply has to go all in.

7.2.2 Scientometrics and Whisper

Scientometrics is the study of the practice of science [147]. Most often, practitioners explore the citations and co-authorships of published research, as well as some content analysis of the publications, to examine the impact of publications, as well as describe how research in a field changes over time. These practitioners would benefit greatly from access to the data set of a fully functional collaborative academic work environment such as Whisper. Most of the analyses performed in the Scientometrics field is done using citation analysis of the bibliographic data associated with a paper, but and only on published papers; the researchers rarely have the opportunity to go back and follow the work that goes in to the development of a paper.

Academic papers are generally limited to a certain page count for publication, which by necessity, also limits the number of citations that can be included; this inherently limits the amount of data available to a Scientometric analysis. A collaboration environment, such as Whisper, would provide a wealth of additional data that might be used in impact analysis. For example, in fields such as biology researchers occasionally publish “method papers”, which describe a new technique (such as a non-invasive means of determining the gender of a lizard [46]); such papers may not have a high citation rate because they don’t introduce new theory. However, if a particular method paper

were found in the digital library for a large number of Projects, even if none of the publications from that Project cited it, that would suggest the paper should have a greater impact value than its sheer citation count would indicate.

7.2.3 Scientometrics and Twitter

There has been some Scientometric-related research within the Twitter social network. For example, Priem and Costello investigated whether or not academics include citations in their tweets [136], while Eysenbach examined how well tweets about a paper immediately after its publication predicted future citation counts [59]. But there are still areas that can be explored.

One area that is of particular interest is the comparison of “sleeping beauties” in scientific literature and in Twitter. In Scientometrics, a sleeping beauty is a paper that is initially cited poorly (or not at all), and is eventually discovered well after the standard window in which publications attract attention has closed [72, 160]. While tweets might have a shorter window in which they would generally be retweeted [115] than scientific papers would generally be cited, that does not preclude their discovery and viral spread at a later date.

Gao and Guan recently explored how knowledge diffusion could be measured through the network of researchers [68]. Their work builds on epidemic models of information diffusion developed within the Scientometric community by Bettencourt et al. [14, 15] and Kiss et al. [103] (which is similar to the methods employed to examine news diffusion in social networks described by Lerman and Ghosh [115]). Their time-

based diffusion analysis treated citations as social connections and was able to identify when a particular researcher's exposure to the topic reached its tipping point (leading to publication on related to that topic). Their analysis is also able to identify key papers related to the topic being examined.

The analysis presented by Gao and Guan is compelling, but it is limited by only having access to published papers. While one could assume every person was exposed to a topic upon publication, that is not always the case. Some might only notice the paper a few months or a year later, when it is cited in a paper that did catch their eye. Additionally, there are many social connections within a community of research that are not captured by citations, so some exposure may happen through colleagues rather than the paper being cited. With access to the daily social connections within a Whisper system, and being able to see both when a paper was added to a personal or project digital library (and potentially the person, project, or community that acted as the source for this exposure), a more detailed and fine-grained analysis could be obtained. Such an analysis might be able to identify additional information, such key evangelists in the spread of the topic whose first publication came long after their initial exposure.

7.3 Final Thoughts

This research was inspired by sporadic observations of various inefficient methods for managing the process of academic research used by both graduate students and professors in various disciplines, and the realization that increased focus in one's area of expertise can limit one's awareness of other research and make cross-cutting research

harder to perform. The model of academic knowledge transfer presented here will hopefully provide a useful tool for exploring how academics can improve their own work processes, whether at the individual, group, or community level. The results presented in this thesis provide insight into not only the process of academic knowledge transfer, but also demonstrate methods by that process can be made more efficient. The goals and requirements of the software system provide a solid foundation upon which a complete knowledge sharing system can be built today and will provide a solid basis for future research and adjustment of both the requirements and the underlying model. In a world growing ever more complex, systems that help us manage that complexity will be among our most valuable tools.

Appendix A

Facebook Survey Questions

A.1 Intake Survey Questions

1. What is your gender?
2. How long have you been using Facebook?
3. What other social networking sites do you use? (Check all that apply)
4. How much time do you spend using the internet each day?
5. Are you currently pursuing a degree (post high school or GED)? (If you are on summer break and plan to attend classes this fall, please choose yes)
6. If you are currently a student or affiliated with an educational institution, what school do you attend? (If you are on summer break or working, please enter the school you most recently attended.)
7. If you are currently a student, what is your major? If you are not currently a

student, but have a degree beyond high school, what was your major?

8. If you are not currently a student, what is your profession or industry?
9. How often do you share links to information (such as news or academic articles, papers, websites) using email?
10. How often do you share links to information (such as news or academic articles, papers, websites) using social networking sites (like your Facebook wall or status)?
11. How often do you share links to information (such as news or academic articles, papers, websites) using an instant messenger?
12. How often do you share links to information (such as news or academic articles, papers, websites) by leaving notes or sending paper mail?
13. How often do you share links to information (such as news or academic articles, papers, websites) verbally, either in person or over the phone?
14. How much do you like sharing information using email?
15. How much do you like sharing information using social networking sites (like your Facebook wall or status)?
16. How much do you like sharing information using an instant messenger?
17. How much do you like sharing information by leaving notes or sending paper mail?
18. How much do you like sharing information verbally, either in person or over the phone?

A.2 Final Survey Questions

1. What is your gender?
2. How long have you been using facebook?
3. How often do you log in to facebook?
4. What is the name of the Facebook application you used?
5. How did you find out about the application?
6. How often do you read new papers?
7. How do you read papers? As in, do you print them out, or read them on a computer screen? Maybe you use an eBook reader, or a mobile phone?
8. Where do you make notes when you're reading?
9. How do you normally keep track of your notes?
10. How many times do you think you went to the the app (Tidbitz, Nuggetz, or Read All About It)?
11. How many Tidbits/Nuggets/Articles did you store there?
12. Did you find it easy to use? If not, what was confusing?
13. Under what circumstances would you use the app more?
14. For each of the following statements, please indicate your level of agreement:
Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree

- (a) I have no interest in social networking websites
- (b) I do not like facebook
- (c) I do not have time for facebook
- (d) Facebook is something I use for fun
- (e) I don't want to do work with Facebook
- (f) It is inconvenient to have facebook open while I'm trying to get work done
- (g) I am not sure how to use the application
- (h) I don't see the point of storing links to papers in the application
- (i) I'm not reading many new papers right now.
- (j) When I'm reading papers, I never remember to open the application until later (if at all).
- (k) I can't trust that the application (and my notes within it) will be around long enough for it to be useful to me.
- (l) I don't trust my notes to applications I don't control
- (m) I have a good system in place for managing my notes already.
- (n) My paper/notes management system is too large for me to change it.
- (o) I would change my paper/notes management system if I got clear benefits from the new system.

15. What benefits would you expect, or want to see, in a notes/paper organization system?

Appendix B

List of Celebrity Twitter Users

Table B.1: Technology

Wil Shipley	Dave Thomas	Guy Kawasaki	John C. Dvorak
Robert Scoble	Jyri Engeström	Ward Cunningham	Tim O'Reilly
Caterina Fake	Joe Hewitt	David Heinemeier Hansson	Anil Dash
Joel Spolsky	John Gruber	Steve Wozniak	Om Malik
Rod Johnson	Ümit Yalçınalp	Padmasree Warrior	Biz Stone

Table B.2: Politics

Cory Booker	Russ Feingold	Michele Bachmann	Claire McCaskill
Nick Clegg	Chuck Grassley	Arnold Schwarzenegger	Kevin McCarthy
Orrin Hatch	Eric Cantor	Thaddeus McCotter	Dennis Kucinich
George Miller	Jennifer Granholm	John Kasich	Ron Paul
Jared Polis	Gavin Newsom	Mark Begich	Jim DeMint

Table B.3: Business

Burt Helm	George Colony	Tim Bradshaw	Bill Gross
Steve Case	Maria A. Andros	Doug Ulman	Rick Myers
Norman Hajar	John Jantsch	Kara Swisher	Craig Newmark
Jack Welch	Amy Cosper	John Lilly	John A. Byrne
Emily Steel	Soren Macbeth	Simeon Simeonov	Richard Branson

Table B.4: Entertainment

Felicia Day	Neil Patrick Harris	Martha Stewart	Kevin Spacey
Donnie Wahlberg	Justine Bateman	Kevin Smith	Margaret Cho
Nathan Fillion	MC Hammer	Sarah Silverman	Kirstie Alley
William Shatner	Alyssa Milano	Jimmy Fallon	Fran Drescher
Stephen Fry	Jane Seymour Fonda	Russell Brand	Brent Spiner

Table B.5: Higher Education

Alec Couros	Nouriel Roubini	Laura Nicosia	Bill Genereux
Marcus du Sautoy	Jonathan Becker	Judy O'Connell	Patrick Strother
Dean Terry	Kent Gustavson	Alfred Hermida	Monty Craig
Barbara B. Nixon	Steve Katz	Paul Bradshaw	Martin Weller
Ryan Seslow	Bernie Dodge	Jay Rosen	Ligon Duncan

Bibliography

- [1] Abhik Sen. Twitter finally becomes platform for politicians. <http://indiatoday.intoday.in/story/twitter-becomes-platform-for-politicians/1/150060.html>, September 2011.
- [2] ACM Digital Library. <http://portal.acm.org>.
- [3] Robert M. Akscyn, Donald L. McCracken, and Elise A. Yoder. KMS: a distributed hypermedia system for managing knowledge in organizations. *Communications of the ACM*, pages 820–835, 1988.
- [4] Alex Eichler. Kim Kardashian Bails Out of Twitter for the Children. <http://www.theatlanticwire.com/entertainment/2010/11/kim-kardashian-bails-out-of-twitter-for-the-children/18319/>, November 2010.
- [5] Francisco Alvarez-Cavazos, Roberto Garcia-Sanchez, David Garza-Salazar, Juan C. Lavariega, Lorena G. Gomez, and Martha Sordia. Universal access architecture for digital libraries. In *CASCON '05: Proceedings of the 2005 conference*

- of the Centre for Advanced Studies on Collaborative research, pages 12–28. IBM Press, 2005.
- [6] Paul André, Michael Bernstein, and Kurt Luther. Who gives a tweet?: evaluating microblog content value. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, pages 471–474, New York, NY, USA, 2012. ACM.
- [7] Marc Andreessen. NCSA Mosaic Technical Summary. <ftp://ftp.ncsa.uiuc.edu/Mosaic/Papers/mosaic.ps.Z>.
- [8] Angela West. How Twitter Web Analytics Will Help Your Business. http://www.pcworld.com/businesscenter/article/239981/how_twitter_web_analytics_will_help_your_business.html, September 2011.
- [9] Charles Bazerman. *Shaping Written Thought: The Genre and Activity of the Experimental Article in Science*. The University of Wisconsin Press, 1988.
- [10] W. Lance Bennett and Shanto Iyengar. A new era of minimal effects? the changing foundations of political communication. *Journal of Communication*, 58(4):707–731, 2008.
- [11] Richard Bentley, Wolfgang Appelt, Uwe Busbach, Elke Hinrichs, David Kerr, Klaas Sikkel, Johnathan Trevor, and Gerd Woetzel. Basic support for cooperative work on the world wide web. *International Journal of Human-Computer Studies*, 46(6):827–846, June 1997.

- [12] Richard Bentley, Thilo Horstmann, Klaas Sikkels, Johnathan Trevor, and Gerd Woetzel. Supporting collaborative information sharing with the world wide web: The BSCW shared workspace system. *Proceedings of the 4th International World Wide Web Conference*, pages 827–846, 1995.
- [13] Tim Berners-Lee. Information Management: A Proposal. <http://www.w3.org/History/1989/proposal.html>.
- [14] L. Bettencourt, A. Cintrón-Arias, D.I. Kaiser, and C. Castillo-Chávez. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*, 364:513–536, 2006.
- [15] L.M.A. Bettencourt, D.I. Kaiser, J. Kaur, C. Castillo-Chavez, and D.E. Wojick. Population modeling of the emergence and development of scientific fields. *Scientometrics*, 75(3):495–518, 2008.
- [16] BibTeX. <http://www.bibtex.org>.
- [17] Michael Bieber, Douglas C. Engelbart, Richard Furuta, Starr Roxanne Hiltz, John Noll, Jennifer Preece, Edward A. Stohr, Murray Turoff, and Bartel van de Walle. Toward virtual community knowledge evolution. *Journal of Management Information Systems*, 18(4):11–35, 2002.
- [18] Wiebe E. Bijker. The social construction of bakelite: Toward a theory of invention. In Wiebe E. Bijker, Thomas P. Hughes, and Trevor J. Pinch, editors, *The*

Social Construction of Technological Systems: New Directions in the Sociology and History of Technology, pages 17–50. MIT Press, Cambridge, MA, 1987.

- [19] Blogger. <http://www.blogger.com>.
- [20] G.W. Bock, R.W. Zmud, Y.G. Kim, and J.-N. Lee. Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate. *MIS Quarterly*, 29(1):87–111, 2005.
- [21] Pierre Bourdieu. *Outline of a Theory of Practice*. Cambridge University Press, New York, 1977.
- [22] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. *Hawaii International Conference on System Sciences*, pages 1–10, 2010.
- [23] danah m boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2008.
- [24] Christina Brodersen and Ole Sejer Iversen. ecell: spatial it design for group collaboration in school environments. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, GROUP '05, pages 227–235, New York, NY, USA, 2005. ACM.
- [25] Andrew L. Brooks and Elizabeth F. Churchill. Tune in, tweet on, twit out: Information snacking on twitter. CHI 2010 Workshop on Microblogging, April 2010.

- [26] Keith D. Brouthers, Ram Mudambi, and David M. Reeb. The blockbuster hypothesis: influencing the boundaries of knowledge. *Scientometrics*, 90(3):959–982, March 2012.
- [27] John Seely Brown and Paul Duguid. Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science*, 2(1):40–57, 1991.
- [28] John Seely Brown and Paul Duguid. Organizing knowledge. *California Management Review*, 40(3):90–111, 1998.
- [29] John Seely Brown and Paul Duguid. Knowledge and organization: A social-practice perspective. *Organization Science*, 12(2):198–213, 2001.
- [30] Frank Buschmann, Regine Meunier, Hans Rohnert, Peter Sommerlad, and Michael Stal. *Pattern-Oriented Software Architecture: A System of Patterns*. John Wiley and Sons, 1996.
- [31] Vannevar Bush. As we may think. *The Atlantic Monthly*, pages 101–108, July 1945.
- [32] Angel Cabrera and Elizabeth F. Cabrera. Knowledge-Sharing dilemmas. *Organization Studies*, 23(5):687–710, September 2002.
- [33] Caitlin Dineen. South Jersey politicians turn to social networking to connect with voters. <http://www.pressofatlanticcity.com/news/press/atlantic/>

south-jersey-politicians-turn-to-social-networking-to-connect-with/
article_19345004-dce0-11e0-934f-001cc4c002e0.html, September 2011.

- [34] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, May 2010.
- [35] Shih-Yuarn Chen, Chia-Ning Chang, Ming-Jin Hwang, Hao-Ren Ke, and Wei-Pang Yang. Architecture for personal digital library. In *ICCOMP'05: Proceedings of the 9th WSEAS International Conference on Computers*, pages 1–6, Stevens Point, Wisconsin, USA, 2005. World Scientific and Engineering Academy and Society (WSEAS).
- [36] CiteULike. <http://www.citeulike.org>.
- [37] Geoffrey Clemmm, Jim Amsden, Tim Ellison, Chris Kaler, and E. James Whitehead. RFC 3253: Versioning extensions to WebDAV (Web Distributed Authoring and Versioning).
- [38] Conja Colete, Carina de Villiers, and Sumarie Roodt. Facebook as an academic tool for ict lecturers. In *SACLA '09: Proceedings of the 2009 Annual Conference of the Southern African Computer Lecturers' Association*, pages 16–22, New York, NY, USA, 2009. ACM.

- [39] Jeff Conklin. Designing organizational memory: Preserving intellectual assets in a knowledge economy. <http://cognexus.org/dom.pdf>, 1997.
- [40] Vignette Corporation. Vignette collaboration. Technical report, http://www.vignette.com/Downloads/WP_VignCollaboration.pdf, May 2004.
- [41] Lisa Covi. Social worlds of knowledge work: How researchers appropriate digital libraries for scholarly communication. *The Digital Revolution: Assessing the Impact on Business, Education, and Social Structures, ASIS Mid-Year Meeting*, pages 84–100, 1996.
- [42] Robert T Craig. Communication theory as a field. *Communication Theory*, 9(2):119–161, 1999.
- [43] Michael Curtin, Nathan Carpenter, and Chris Ritzo. Adding fun and games to training programs. In *SIGUCCS '06: Proceedings of the 34th annual ACM SIGUCCS conference on User services*, pages 50–54, New York, NY, USA, 2006. ACM.
- [44] Dave Morin. Announcing Facebook Connect. <https://developers.facebook.com/blog/post/2008/05/09/announcing-facebook-connect/>, May 2008.
- [45] David Heinemeier Hansson. David Heinemeier Hansson’s Twitter feed. <http://twitter.com/dhh>.
- [46] A.R. Davis and D.H. Leavitt. Candlelight vigilis: a noninvasive method for sexing small, sexually monomorphic lizards. *Herpetological Review*, 38(4):402, 2007.

- [47] Camille Bierens de Haan, Gilles Chabré, Francis Lapique, Gil Regev, and Alain Wegmann. Oxyoron, a non-distance knowledge sharing tool for social science students and researchers. *Proceedings of the international ACM SIGGROUP conference on Supporting Group Work*, pages 219–228, November 1999.
- [48] Delicious Library. <http://www.delicious-monster.com>.
- [49] Joan DiMicco, David R. Millen, Werner Geyer, Casey Dugan, Beth Brownholtz, and Michael Muller. Motivations for social networking at work. In *CSCW '08: Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work*, pages 711–720, New York, NY, USA, 2008. ACM.
- [50] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. The benefits of Facebook “friends:” social capital and college students’ use of online social network sites. *Journal of Computer Mediated Communication*, 12(4), 2007.
- [51] Richard M. Emerson. Social exchange theory. *Annual Review of Sociology*, 2(1976):335–362, 1976.
- [52] EndNote. <http://www.endnote.com>.
- [53] Douglas C. Engelbart. Program on human effectiveness. Technical report, Stanford Research Institute, August 1961.
- [54] Douglas C. Engelbart. Special considerations of the individual as a user, generator, and retriever of information. *American Documentation*, 12(2):121–125, April 1961.

- [55] Douglas C. Engelbart. Augmented human intellect program. Technical report, Stanford Research Institute, March 1962.
- [56] Douglas C. Engelbart. Study for the development of human intellect augmentation techniques. Technical report, Stanford Research Institute, July 1968.
- [57] Douglas C. Engelbart. Toward high-performance organizations: A strategic role for groupware. *Proceedings of the Groupware '92 Conference*, August 1992.
- [58] Douglas C. Engelbart, Richard W. Watson, and James C. Norton. The augmented knowledge workshop. *AFIPS Conference Proceedings*, 24, June 1973.
- [59] G. Eysenbach. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4), 2011.
- [60] Study: 54 percent of companies ban facebook, twitter at work.
<http://www.wired.com/epicenter/2009/10/study-54-of-companies-ban-facebook-twitter-a>
October 2009.
- [61] Study: 54 percent of companies ban facebook, twitter at work. <http://www.wired.com/epicenter/2009/10/study-54-of-companies-ban-facebook-twitter-at-work/>, October 2009.
- [62] Facebook platform. <http://developer.facebook.com>.
- [63] Lourdes Fernández, J. Alfredo Sánchez, and Alberto García. Mibiblio: personal

- spaces in a digital library universe. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 232–233, New York, NY, USA, 2000. ACM.
- [64] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk Nielsen, Larry Masinter, Paul Leach, and Tim Breners-Lee. RFC 2616: Hypertext transfer protocol — HTTP/1.1.
- [65] Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.
- [66] Martin Fishbein and Icek Ajzen. *Belief, attitude, intention and behaviour: An introduction to theory and research*. Addison-Wesley, 1975.
- [67] Batya Friedman, Peter H. Kahn, and Alan Borning. Value sensitive design and information systems. In *Human-Computer Interaction and Management Information Systems: Foundations. M.E. Sharpe*, pages 348–372, 2006.
- [68] Xia Gao and Jiancheng Guan. Network model of knowledge diffusion. *Scientometrics*, 90(3):749–762, March 2012.
- [69] L. Nancy Garrett, Karen E. Smith, and Norman Meyrowitz. Intermedia: issues, strategies, and tactics in the design of a hypermedia document system. *Proceedings of the 1986 ACM conference on Computer-supported cooperative work*, 1986.
- [70] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong. Mylifebits: fulfilling the memex vision. In *MULTIMEDIA '02: Proceedings of the*

- tenth *ACM international conference on Multimedia*, pages 235–238, New York, NY, USA, 2002. ACM.
- [71] David Gingell. A fifteen minute guide to enterprise content management. Technical report, Documentum, 2003.
 - [72] W Glanzel, B Schlemmer, and B Thijs. Better late than never? on the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58(3):571–586, 2003.
 - [73] Jeremy Goecks and Elizabeth D. Mynatt. Leveraging social networks for information sharing. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, CSCW '04, pages 328–331, New York, NY, USA, 2004. ACM.
 - [74] Yaron Goland, E. James Whitehead, Asad Faizi, Steve Carter, and Del Jensen. RFC 2518: HTTP extensions for distributed authoring — WebDAV.
 - [75] Marcos Gonçalves, Edward Fox, Layne Watson, and Neill A. Kipp. Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. *Transactions on Information Systems*, 2(2):270–312, April 2004.
 - [76] Danny Goodman. *The Complete HyperCard Handbook*. Bantam Books, 1987.
 - [77] M. S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.
 - [78] Martin Gudgin, Marc Hadley, Noah Mendelsohn, Jean-Jacque Moreau, and Hen-

- rik Frystyk Nielsen. Simple Object Access Protocol. <http://www.w3.org/TR/soap12-part1/>.
- [79] Morten T. Hansen. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*, 44(1):82–111, 1999.
- [80] Morten T Hansen. Knowledge networks: Explaining effective knowledge sharing in multiunit companies. *Organization Science*, 13(3):232–248, 2002.
- [81] Brian Heater. Look Out, Twitter: Here Comes Oprah. <http://www.pcmag.com/article2/0,2817,2345462,00.asp>, April 2009.
- [82] Herman F. Hegner. Scientific value of the social settlements. *American Journal of Sociology*, 3(2):171–182, 1897.
- [83] Anne Hewitt and Andrea Forte. Crossing boundaries: Identity management and student/faculty relationships on the Facebook. *Poster/Extended Abstract, CSCW 2006*, 2006.
- [84] David L. Hicks and Klaus Tochtermann. Personal digital libraries and knowledge management. *Journal of Universal Computer Science*, 7(7):550–565, jul 2001. http://www.jucs.org/jucs_7_7/personal_digital_libraries_and.
- [85] Mark Horton and Rick Adams. RFC 1036 : Standard for interchange of USENET messages.

- [86] Chin-Lung Hsu and Judy Chuan-Chuan Lin. Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. *Inf. Manage.*, 45(1):65–74, January 2008.
- [87] Bernardo Huberman, Daniel M Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), January 2009.
- [88] Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong regularities in World Wide Web surfing. *Science*, 280(5360):95–97, 1998.
- [89] Lee Humphreys. Historicizing microblogging. CHI 2010 Workshop on Microblogging, April 2010.
- [90] International Business Times. Osama Bin Laden Death: Celebrities tweet reaction. <http://www.ibtimes.com/articles/140375/20110502/osama-bin-laden-celebrities-twitter-tweet-reaction.htm>, May 2011.
- [91] Shanto Iyengar. *Is anyone responsible? How television frames political issues*. University Of Chicago Press, Chicago, IL, 1991.
- [92] Jackrabbit. <http://jackrabbit.apache.org>.
- [93] Jane Jacobs. *The Death and Life of Great American Cities*. Random House, 1961.
- [94] William C. Janssen and Kris Popat. Uplib: a universal personal digital library system. In *DocEng '03: Proceedings of the 2003 ACM symposium on Document engineering*, pages 234–242, New York, NY, USA, 2003. ACM.

- [95] Mohammad Hossein Jarrahi and Steve Sawyer. Social networking technologies and organizational knowledge sharing as a sociotechnical ecology. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, CSCW '12, pages 99–102, New York, NY, USA, 2012. ACM.
- [96] Jeff Joerres. Jeff Joerres’s Twitter feed. <http://twitter.com/ManpowerGroupJJ>.
- [97] Philip Johnson. Supporting exploratory cscw with the egret framework. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, CSCW '92, pages 298–305, New York, NY, USA, 1992. ACM.
- [98] JSR 170: Content Repository for Java technology API. <http://jcp.org/en/jsr/detail?id=170>.
- [99] Atreyi Kankanhalli, Bernard C. Y. Tan, and Kwok kee Wei. Contributing knowledge to electronic knowledge repositories: An empirical investigation. *MIS Quarterly*, 29:113–143, 2005.
- [100] Elihu Katz and Paul F. Lazarsfeld. *Personal Influence*. The Free Press, 1955.
- [101] Alexy Khrabrov and George Cybenko. Discovering influence in communication networks using dynamic graph analysis. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 288–294, Washington, DC, USA, 2010. IEEE Computer Society.
- [102] Tim Kindberg, Nick Bryan-Kinns, and Ranjit Makwana. Supporting the shared care of diabetic patients. In *Proceedings of the international ACM SIGGROUP*

- conference on Supporting group work*, GROUP '99, pages 91–100, New York, NY, USA, 1999. ACM.
- [103] I.Z. Kiss, M. Broom, P.G. Craze, and I. Rafols. Can epidemic models describe the diffusion of topics across disciplines? *Journal of Informetrics*, 4(1):74–82, 2010.
 - [104] Kate Klonick. Facebook 'Feeds' Online Privacy Debate. <http://abcnews.go.com/Technology/story?id=2409970&page=1#.T4t5hZgpzwM>.
 - [105] Peter Kollock. Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, 24(1):183–214, 1998.
 - [106] William H. Kruskal and W. Allen Wallis. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.
 - [107] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.
 - [108] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. The open archives initiative protocol for metadata harvesting. <http://www.openarchives.org/OAI/openarchivesprotocol.html>.
 - [109] Cliff Lampe, Nicole Ellison, and Charles Steinfield. A face(book) in the crowd: social searching vs. social browsing. In *Proceedings of the 2006 20th anniversary*

- conference on Computer supported cooperative work*, CSCW '06, pages 167–170, New York, NY, USA, 2006. ACM.
- [110] Cliff Lampe, Nicole Ellison, and Charles Steinfield. A Face(book) in the crowd: Social searching vs. social browsing. In *CSCW '06: Proceedings of the ACM 2006 on Computer Supported Cooperative Work*, pages 167–170, New York, NY, USA, 2006. ACM.
- [111] Harold D. Lasswell. The structure and function of communication in society. In Lyman Bryson, editor, *The Communication of Ideas*, pages 37–51. Institute for Religious and Social Studies, New York, NY, 1948.
- [112] Amy Law and Raylene Charron. Effects of agile practices on social factors. In *Proceedings of the 2005 workshop on Human and social factors of software engineering*, HSSE '05, pages 1–5, New York, NY, USA, 2005. ACM.
- [113] Paul F. Lazarsfeld, Bernard Berelson, and Hazel Gaudet. *The People's Choice*. Columbia University Press, third edition, 1968.
- [114] Jennifer LeClaire. Report: Facebook's Privacy Promises Flawed. http://www.newsfactor.com/story.xhtml?story_id=57013.
- [115] Kristina Lerman and Rumi Ghosh. Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, March 2010.

- [116] Bo Leuf and Ward Cunningham. Wiki: Wat is wiki. Technical report, Wiki.org, June 2002.
- [117] Meira Levy and Orit Hazzan. Knowledge management in practice: The case of agile software development. In *Proceedings of the 2009 ICSE Workshop on Cooperative and Human Aspects on Software Engineering*, CHASE '09, pages 60–65, Washington, DC, USA, 2009. IEEE Computer Society.
- [118] J. C. R. Licklider. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1:4–11, March 1960.
- [119] Linkedin. <http://www.linkedin.com>.
- [120] Lucene. <http://lucene.apache.org>.
- [121] Yi Ma, Edward A. Fox, and Marcos A. Gonçalves. Personal digital library: pim through a 5s perspective. In *PIKM '07: Proceedings of the ACM first Ph.D. workshop in CIKM*, pages 117–124, New York, NY, USA, 2007. ACM.
- [122] Richard MacManus. New Facebook Privacy Options Go Live - May Overwhelm Users. http://www.readwriteweb.com/archives/new_facebook_privacy_options_go_live.php.
- [123] Agile Manifesto. Manifesto for Agile Software Development. <http://agilemanifesto.org/>.
- [124] Frank J Massey. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

- [125] Peter McCullagh and John A. Nelder. *Generalized linear models (Second edition)*. Chapman & Hall, New York, NY, 1989.
- [126] Mendeley. <http://www.mendeley.com>.
- [127] Jessica K. Miller, Batya Friedman, and Gavin Jancke. Value tensions in design: the value sensitive design, development, and appropriation of a corporation’s groupware system. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP ’07, pages 281–290, New York, NY, USA, 2007. ACM.
- [128] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW ’10, pages 189–192, New York, NY, USA, 2010. ACM.
- [129] Janine Nahapiet and Sumantra Ghoshal. Social Capital, Intellectual Capital, and the Organizational Advantage. *The Academy of Management Review*, 23(2):242–266, 1998.
- [130] The national science digital library. <http://www.ndsl.org>.
- [131] Gerard Oleksik, Natasa Milic-Frayling, and Rachel Jones. Beyond data sharing: artifact ecology of a collaborative nanophotonics research centre. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW ’12, pages 1165–1174, New York, NY, USA, 2012. ACM.
- [132] Karl Pearson. On the criterion that a given system of deviations from the probable

- in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.
- [133] Trevor J. Pinch and Wiebe E. Bijker. The social construction of facts and artifacts: Or how the sociology of science and the sociology of technology might benefit each other. In Wiebe E. Bijker, Thomas P. Hughes, and Trevor J. Pinch, editors, *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*, pages 17–50. MIT Press, Cambridge, MA, 1987.
- [134] Alan L. Porter and Ismael Rafols. Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, 81(3):719–745, December 2009.
- [135] PostgreSQL. <http://www.postgresql.org>.
- [136] J. Priem and K.L. Costello. How and why scholars cite on twitter. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, 2010.
- [137] Maryam N. Razavi and Lee Iverson. A grounded theory of information sharing behavior in a personal learning space. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, CSCW '06*, pages 459–468, New York, NY, USA, 2006. ACM.
- [138] Byron Reeves and J. Leighton Read. *Total Engagement: Using Games and Vir-*

- tual Worlds to Change the Way People Work and Businesses Compete*. Harvard Business School Press, 2009.
- [139] Natalia Reyes-Farfán and J. Alfredo Sánchez. Personal spaces in the context of oai. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 182–183, Washington, DC, USA, 2003. IEEE Computer Society.
- [140] Robert Kiyosaki. Robert Kiyosaki’s Twitter feed. <http://twitter.com/theRealKiyosaki>.
- [141] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In *Proceedings of the ECML/PKDD 2011*, 2011.
- [142] Sarah Michael. Savvy business owners use Twitter and Facebook to nab new customers and drive up sales. <http://www.heraldsun.com.au/business/entrepreneur/small-business-turns-to-the-internet/story-fn7ve51s-1226130664356>, September 2011.
- [143] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 3(52), 1965.
- [144] Shea Bennett. President Obama Becomes Third Human, First Politician To Reach 10 Million Twitter Followers. http://www.mediabistro.com/alltwitter/barack-obama-twitter-10-million-followers_b13599, September 2011.
- [145] Dawei Shen, Marshall Van Alstyne, Andrew Lippman, and Hind Benbya. Barter:

- mechanism design for a market incented wisdom exchange. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, pages 275–284, New York, NY, USA, 2012. ACM.
- [146] Meredith M. Skeels and Jonathan Grudin. When social networks cross boundaries: a case study of workplace use of Facebook and LinkedIn. In *Proceedings of the ACM 2009 international conference on Supporting group work, GROUP '09*, pages 95–104, New York, NY, USA, 2009. ACM.
- [147] Derek R. Smith. Impact factors, scientometrics and the history of citation-based research. *Scientometrics*, 92(2, SI):419–427, August 2012.
- [148] Spring Framework. <http://springframework.org>.
- [149] Sushmita Subramanian and Wendy March. Sharing presence: Can and should your tweets be automated? CHI 2010 Workshop on Microblogging, April 2010.
- [150] Subversion. <http://subversion.tigris.org>.
- [151] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proceedings of the IEEE Second International Conference on Social Computing (Social-Com)*, pages 177–184, August 2010.
- [152] Yasuyuki Sumi and Kenji Mase. Supporting awareness of shared interests and experiences in community. *SIGGROUP Bull.*, 21(3):35–42, December 2000.
- [153] Tony Hsieh. Tony Hsieh’s Twitter feed. <http://twitter.com/zappos>.

- [154] Jonathan Trevor, Tom Rodden, and John Mariani. The use of adapters to support cooperative sharing. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, CSCW '94, pages 219–230, New York, NY, USA, 1994. ACM.
- [155] Turning work into play with online games. <http://hplusmagazine.com/articles/art-entertainment/turning-work-play-online-games>, January 2010.
- [156] How Tweet It Is!: Library Acquires Entire Twitter Archive. <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/>.
- [157] Twitter. <http://twitter.com>.
- [158] Twitter api. <http://dev.twitter.com/>.
- [159] Statistical summary of students and staff. <http://www.ucop.edu/ucophome/uwnews/stat/statsum/fall2008/statsumm2008.pdf>, 2008.
- [160] AFJ van Raan. Sleeping Beauties in science. *Scientometrics*, 59(3):467–472, 2004.
- [161] Mitch Wagner. Turning work into play is no game. http://www.informationweek.com/blog/main/archives/2008/02/turning_work_in.html, February 2008.
- [162] Joseph B. Walther, Caleb T. Carr, Scott Seung W. Choi, David C. DeAndrea, Jinsuk Kim, Stephanie Tom Tong, and Brandon Van Der Heide. Interaction of in-

- terpersonal, peer, and media influence sources online. In Zizi Papacharissi, editor, *A Networked Self*. Routledge, New York, NY, 2011.
- [163] Dadong Wan and Philip M. Johnson. Computer supported collaborative learning using clare: the approach and experimental findings. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work, CSCW '94*, pages 187–198, New York, NY, USA, 1994. ACM.
- [164] Sheng Wang and Raymond A Noe. Knowledge sharing: A review and directions for future research. *Human Resource Management Review*, 20(2):115–131, 2010.
- [165] M M Wasko and S Faraj. Why should i share? examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly*, 29(1):35–57, 2005.
- [166] Omar Wasow, Alex Baron, Marlon Gerra, Katharine Lauderdale, and Han Zhang. Can tweets kill a movie? an empirical evaluation of the bruno effect. CHI 2010 Workshop on Microblogging, April 2010.
- [167] Jane Webster and Joseph J. Martocchio. Turning work into play: Implications for microcomputer software training. *Journal of Management*, 19(1):127–146, 1993.
- [168] G Widen-Wulff. Explaining knowledge sharing in organizations through the dimensions of social capital. *Journal of Information Science*, 30(5):448–458, 2004.
- [169] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

- [170] Fang Wu, Bernardo A. Huberman, Lada A. Adamic, and Joshua R. Tyler. Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337(1-2):327–335, 2004.
- [171] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW ’11, pages 705–714, New York, NY, USA, 2011. ACM.
- [172] Wei-Li Wu, Bi-Fen Hsu, and Ryh-Song Yeh. Fostering the determinants of knowledge transfer: a team-level analysis. *J. Inf. Sci.*, 33(3):326–339, June 2007.
- [173] XFire. <http://xfire.codehaus.org>.
- [174] Toshio Yamagishi and Karen S. Cook. Generalized exchange and social dilemmas. *Social Psychology Quarterly*, 56(4):235–248, 1993.
- [175] S. Ye and S. F. Wu. Measuring Message Propagation and Social Influence on Twitter. com. In *Social Informatics: Second International Conference, SocInfo 2010, Laxenburg, Austria, October 27-29, 2010, Proceedings*, page 216, 2010.
- [176] John R. Zaller. *The Nature and Origin of Mass Opinion*. Cambridge University Press, Cambridge, New York, Oakleigh, 1992.
- [177] Dan Zarrella. The science of retweets. <http://danzarrella.com/the-science-of-retweets-report.html>, September 2009.
- [178] Dejin Zhao and Mary Beth Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the*

ACM 2009 international conference on Supporting group work, GROUP '09, pages
243–252, New York, NY, USA, 2009. ACM.