FREE OR OPEN ACCESS TO SCHOLARLY DOCUMENTATION: GOOGLE SCHOLAR OR ACADEMIC LIBRARIES

A Dissertation
presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by
C. SEAN BURNS
Dr. John M. Budd, Dissertation Supervisor

MAY 2013

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

FREE OR OPEN ACCESS
TO SCHOLARLY DOCUMENTATION:
GOOGLE SCHOLAR OR ACADEMIC LIBRARIES

presented by C. Sean Burns,

a candidate for the degree of doctor of philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor John M. Budd

Professor Denice Adkins

Professor Jenny Bossaller

Professor Heather Moulaison

Professor Alejandro Morales

DEDICATION

I dedicate this dissertation to my wife and our son: Amy and Søren. Without Amy's support and patience, this dissertation would not have been possible. Without their love, it would not have been worth doing. You two mean everything to me. I hope I always make you feel loved.

To my Mom. The strongest person I have ever known. You have been a rock for me, and I owe you so much. To my brother and sister, Scott and Amber. I love you two more than you can imagine. You are amazing, as each of your children are.

To my grandparents, Jo and Papa. You are giants. I miss you so much, and you still inspire me.

To all my crazy, beautiful, extended family. You really are crazy and beautiful, and I love and am thrilled by you all.

And to Tory. I miss you, little girl.

ACKNOWLEDGEMENTS

I have been graced with so much support.

John M. Budd has been an incredible inspiration and a wonderful supervisor, mentor, teacher, and collaborator. I will dearly miss not being able to visit his office unannounced or hunt him down at one of the local coffee shops to share some discussion or to receive some sage advice. "Do as I say and not as I do," he often says, but since he is a scholar, a gentleman, and a profound intellectual, that is the only piece of advice of his that I will ignore.

My thanks to my committee. Dr. Denice Adkins possesses the ability to ask the most fundamental questions. Dr. Heather Moulaison has the most critical eye. Dr. Jenny Bossaller has the most global perspective. Dr. Alejandro Morales, my outside member, is an inspiring, caring, and challenging teacher as well as a model researcher. These four souls have made me a better person. I offer my deepest gratitude for their input, comments, questions, discussions, guidance, and support. If I am successful in my career, much of it will be to their credit.

As an example of wisdom and rigor, Dr. Sanda Erdelez has been a great intellectual influence on me. I will always cherish having been a student of hers and I am grateful for all the cheering she has done for me. Dr. Linda Esser showed me needed kindness when I experienced an injury just before the beginning of the doctoral program. Her strength, her caring, and her work ethic motivate me.

My thanks to Dr. Thomas Kochtanek and to the director of the Pictures of the Year International program, Rick Shaw. Their support in the early years of the program was without bounds.

I wish all the best to my fellow Mizzou travelers, past and present, and would like to thank them for their friendship and for their conversations: Kristine (Tine) Stewart, Ben Richardson, Dr. Xin Wang, Carol Smith, Angela Sample, Dr. Dorian Lange, Dr. Jennifer Crispin, Dr. Said Alghenaimi, Dr. Musa Wakhungu, Nathan Lowrance, Pornsuree Jamsri, Francis Kabaru, Dr. Brooke Shannon, Josipa Basic, Dinara Saparova, and Dr. Borchuluun Yadamsuren. Additionally, Dr. Heather Hill is awesome and makes me want to do awesome things.

James Borwick and Corrie Hutchinson are two of the smartest and most talented people I know. My thanks to them for their endless discussions and thoughtful advice. I will certainly miss the many invigorating debates with Corrie, during breaks from cataloging books at Stephens College, as well as Jim's insightful observations about everything.

I would not be half the person I am today without my undergraduate professors at Monmouth College. My endless thanks to Drs. Robert Cathey, Chenyang Li, David Suda, Gary Wilhardt, Jerry McNamara, and Craig Watson. I could not have had a better undergraduate education. I still draw from my class notes and model my teaching after them.

To three of the most important people from the early part of my academic life: Ms. Billie Bailey, Mrs. Janet Archer, and Mrs. Carr-Anderson. You three are always with me. You made high school a great experience and you made learning a great joy.

To my cherished friends, who help complete this thing we call life: Scott Hodges, Matt Altman, Rhett Merz, Courtney Modesitt, Brandon Oberg, and Shannon Wortham. Thanks also to Nicholas Totten, our roommate and friend, who politely endured many 1

A.M. monologues about academic libraries and scholarly communication.

To Toni Milstead, Amy Adam, Kylee Soucie, and Diane Vacca for their administrative support and their unending kindness. Thank you so much for making life easier.

I would like to thank the librarians and staff at MU Libraries. They have always been friendly and available. Also, I spent a considerable amount of time at the Archives of the University of Missouri. Although my research there was for a separate project, the work informed my perspective on the history of academic libraries, which is important to this project. I would like to thank the archivists, Kris Anstine and Gary Cox, for their generous time and guidance.

I would like to thank CiteULike.org. Their site partially inspired this research and I am gracious for having access to their data.

To the library and information science community. They have been most welcoming and encouraging. I would especially like to express my deepest gratitude to *Beta Phi Mu* and to Dr. Eugene Garfield for the generous Eugene Garfield Doctoral Dissertation Fellowship.

Peace and long-life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Soon after the university movement started in the late 1800s, academic libraries became the dominant providers of the tools and services required to locate and access scholarly information. However, with the advent of alternate discovery services, such as Google Scholar, in conjunction with open access scholarly content, researchers now have the option to bypass academic libraries when they search for and retrieve scholarly information. This state of affairs implies that academic libraries exist in competition with these alternate services and with the patrons who use them, and as result, may be disintermediated from the scholarly information seeking and retrieval process.

Drawing from decision and game theory, bounded rationality, information seeking theory, citation theory, and social computing theory, this dissertation uses bibliometrics to investigate the state of affairs. The purpose is to understand if and how academic librarians are responding as competitors to changing scholarly information seeking and collecting practices. Bibliographic data was collected in 2010 from a systematic random sample of references on CiteULike.org and analyzed with three years of bibliometric data collected from Google Scholar. Findings indicate that scholars collect articles that can be located and retrieved without the benefit of a university's proxy and with services like Google Scholar. Although this suggests that academic libraries are being disintermediated, an analysis of the sources providing access indicates that academic libraries are key providers of free and open access content through a number of venues, including institutional repositories. These findings suggest that academic librarians are playing competitively.

1 INTRODUCTION

1.1 Statement of the Problem

In 2010, Ithaka S+R published the results of a 2009 survey which asked faculty about their scholarly communication behaviors and attitudes. The survey gives some credence to the following key observation:

> Basic scholarly information use practices have shifted rapidly in recent years, and as a result the academic library is increasingly being disintermediated from the discovery process, risking irrelevance in one of its core functional areas (Schonfeld & Housewright, 2010, p. 2).

Contrary to recent studies that suggest increased usage of the academic library (e.g., Budd, 2009), the report suggests that researchers in the sciences, social sciences, and the humanities have moved away from the library building, the librarians, and the library's catalog and databases and have moved towards general purpose search engines and other electronic resources to find and satisfy their document needs. Although search and discovery through electronic services include those subscribed by the library, the report reveals, at the network level, the heavy use of non-library electronic discovery services. For instance, searching with Google ranks third in the discovery process (~70%), behind searching electronic, full text databases (~90%), and following citations (~90%), a process some in our field refer to as *chaining* (Ellis, Cox, & Hall, 1993). While only 8.6% out of 35,184 faculty who received the survey responded, and although some have argued that the survey is based on incomplete premises (Nyquist, 2010), the findings warrant additional research about either the central or marginalized role academic libraries function for today's scholars. Thus, this report informs the first research question:

> *RQ 1: Is the current state of affairs, at the network level, such that non-library electronic discovery services marginalize academic libraries?*

The state of affairs at the network level may encourage alternate paths to information, but there might be an additional issue involving open access content. Broadly speaking, open access content is freely accessible to readers with means to the Internet. This is unlike other electronic, scholarly content behind subscription barriers, which require both access to the Internet and access to a library's services (or those provided by other research organizations or with funds to purchase the content directly). Given that open access content is accessible outside a library portal and its collections, if researchers increasingly use non-library electronic discovery services, then non-library electronic discovery services plus the growing availability of open access content make it possible to bypass both the library's services and electronic collections.

Research about the influence and reach of open access content is growing. With its perceived importance for academic libraries, as a publishing model that librarians hope will counteract the growing and unsustainable costs of serials, such influence and reach require examination and inform the second research question:

> *RQ 2: Does open access content, in conjunction with non-library electronic*

discovery services, marginalize academic libraries?

To answer these questions, I use decision and game theory to frame an analysis of a systematic random sample of bibliographic references collected by researchers on CiteULike.org, a social computing bibliographic reference management site. Using these references' bibliometric data, collected from Google Scholar, the main objective is to identify where and how these users have collected these references. Using logistic regression, the second objective is to determine what factors predict or explain collecting these references. Finally, using Bayes' Theorem, the third objective is to build a hypothetical probability profile that illustrates the likelihood that a library's collections are being used given the use of other documents that may be sourced at other locations, such as those held in subject or institutional repositories and which may be found through a service such as Google Scholar. This process allows a determination of whether using non-library discovery services to retrieve open access or freely available content is a relevant alternative to using the library's services to retrieve subscribed content. If the relevant alternative is viable, then the process allows for a determination to be made about the competitiveness of the alternative. Where I define *non-library discovery services, alternate discovery services, relevant alternative, or third party discovery services* in reference to what Ithaka S+R (Schonfeld & Housewright, 2010) describe as “A general purpose search engine on The Internet or World Wide Web such as Google or Yahoo” (p. 4), I can state these allowances as two hypotheses:

H_1: Using a third party discovery service to retrieve open access or freely available content is a relevant alternative to using the library's services to retrieve subscribed content.

H_2: The relevant alternative is a competitive alternative; that is, the relevant alternative entails an outcome where the payoffs are greater than the decision to use the academic library's services and subscribed content.

1.2 Statement of Purpose

The overall goal of this project is to understand the implications that researchers' information seeking and collecting actions have on academic libraries. My hope is the analysis will help academic librarians and library and information science researchers continue to devise strategies that serve their communities' needs given a world where users have many search strategies (Bates, 1981) or choices for searching and retrieving information.

1.3 Significance of the Issue

The impact of open access content and alternate discovery services on the academic library's core function and purpose is part of the significance of this issue, but the main significance is the underlying reason why the impact exists. It is clear that for the first time in history, researchers and other types of library and non-library users have many options available to them to both search for and retrieve good quality information. While the existence of non-library options is non-trivial, what gives the entire search and source domain its real value lies with how and why people make decisions or accomplish their information tasks. The information needs of the user are not met simply by providing relevant collections but by also addressing their decision matrices and by developing an understanding of how these decision matrices might be rational. An introduction to these decision issues is presented in the following section.

1.3.1 Preferences, Utility, Risk, and Prior Information

The Ithaka S+R report (Schonfeld & Housewright, 2010) reveals something about the preferences of information seekers and users. In general, library and information science research has excelled in identifying the preferences of those engaged in information seeking and use. These preferences are often used to help both librarians and information seekers acquire more skills at handling the complex information and knowledge systems that our society is built upon (Julien & Genuis, 2011). However, a list of user preferences can also be applied by librarians to devise appropriate strategies that respond to users' information seeking related actions (e.g., Mullen & Hartman, 2006). In this sense, the preferences that library and information science research have identified serve as a rich source of information for devising and responding to what users want or need in terms of information services and sources but also in terms of organizational needs (see e.g., Theng & Sin, S-C. J. 2012).

Decision and game theory use preferences to rank the payoffs one would expect to receive by applying a decision or strategy (Dixit & Skeath, 2004). The theories help either to explain or to prescribe courses of action either for single individuals or agents or between two or more people or agents whose decisions take into consideration the other's. For example, given an agent's preference to act in a certain way, such as a tacit preference to acquire as much as possible or as much as is needed for as little cost as possible, decision theory provides an analytic framework that describes how an agent makes a decision among a set of relevant alternatives. In the context of this study, the decision may involve the use of a library's or a non-library's search service as a research starting point. Game theory describes how an agent selects a strategy in response to an opposing player's strategy selection. For example, given a user's preference for little effort and much gain, it could be asked what is a librarian's best strategic response. In this research, I take the abstract view that librarians function as one player and researchers, as information seekers and users (in general), function as an opposing player. This relationship is motivated by a simple explanatory heuristic (Abbott, 2004), which places front and center the notion that a strategic interaction exists between librarians and members of their communities, since the former attempt to offer the best search and retrieval services and the latter attempt to satisfy their search efforts using whatever relevant search services are available to them.

George Kingsley Zipf (1949) termed the *principle of least effort* to describe what he derived as a natural tendency among individuals not simply to minimize their work but their *probable average rate of work*. He used the phrase *principle of least effort* to describe this tendency but in doing so, the focus on the *probable* aspect of the principle sometimes gets lost, even though this may be the most important part. In emphasizing the *probable*, it becomes more apparent that our actions to minimize our probable average rate of work are based on the information we have regarding those probabilities or, lacking complete information, the predictive expectations (Nickel, 2009) or beliefs we have about them. Consequently, even if we intend to minimize our probable average rate of work, we may or may not be successful given what we know, expect, trust, or believe will do so. This implies that given two decisions concerning where to start one's research, even if one starting point is more likely to minimize our probable average rate of work over the long run, if we do not expect it to then we may not, on average, select that option (c.f., Savolainen, 2012).

Although Zipf describes the principle of least effort as a natural human characteristic or behavior of individuals, within the framework used in this study I take the view that the principle of least effort can be described as a *preference of least effort*. The semantic substitution simply places more emphasis on the notion that what explains our tendencies and choices are often varied (Hausman, 2005) and actionable, in that we are able to act on our preferences. Despite the terminology, we might posit that some choose Google Scholar as a research starting point because their preferences for locating information include maximizing their success for finding information while minimizing their probable effort to do so, and let us suppose they believe Google Scholar is good at this. At the same time, some may choose the library or some aspect of it, such as its web site, as a research starting point because their same preference for least effort involves the library as a starting point, and they believe using the library minimizes their probable effort in locating relevant information.

The important question for librarians concerns what users are doing in the aggregate. If researchers tend to select a third party search service as a research starting point as often as a library's search service (e.g., Niu et al., 2010), then it may not be because these researchers believe that the library's services cannot satisfy their information needs;

rather, it may very well be because these researchers believe that using the library's search service means greater effort, or greater cost, given both the possible outcomes or payoffs and their other options. The question then is how much of a payoff does one need to pursue a decision when it is believed to be costly? Or what incentives are needed to encourage maximizing and not just satisficing (Simon, 1955), where to maximize indicates acquiring the highest possible payoff? Or, alternatively, how can the use of an academic library, or the conscious decision to choose the academic library as a research starting point, be viewed or believed to be a satisficing function and not a maximizing function? These alternate choices are always in opposition to the other; hence, while the principle of least effort is an interesting concept alone, it is especially interesting when it is placed alongside relevant alternatives. When valuating a library's services or its collections in order to determine, for example, a return on investment (e.g., Tenopir, 2012), that value cannot be determined in isolation from the value of a relevant alternative, just as the value of real estate cannot be determined without tracking adjacent property values (e.g., Farber, 1998). Thus, for example, we could ask what is the academic library's value given the existence of a thing like Google Scholar?

Let us consider a hypothetical. If someone guarantees me $10 to perform a task involving minimal effort or $20 to perform a task involving greater effort, which task will I select? This depends on several factors. One, it depends on my current need and wealth (Brandstätter & Brandstätter, 1996). If I have no wealth and am trying to determine how to purchase my next meal, perhaps it is likely that I will choose the more difficult task for $20 in order to increase my payoff. However, if I have a few hundred dollars in hand, the *law of diminishing returns* suggests that it is likely I will choose the easy task since the difference between $20 and $10 is less important to me.

The subjective utility of either task may also depend on my risk attitude (Rabin, 2000). Let us stipulate that the payout is guaranteed only if I succeed in the task, and let us define the minimal and maximal efforts by the probability of successful completion. Now there is a certain risk associated with the success of the outcome and earning the payoff. This risk might involve my belief about whether I can accomplish the task. Consequently, I believe the task that involves minimal effort will be less risky with a probability of success at 0.70 and I believe the task that involves greater effort will be riskier with a probability of success at 0.30. It is important to note that, in this case with few qualifications, only the risk-seeking person chooses the path of greater effort. Both the risk-averse and the risk-neutral persons will prefer the path of least effort (see also Tversky & Kahnaman, 1974; Kahneman & Tversky, 1979).

A third factor involves prior information (Schmeidler, 1989). Here I illustrate this by way of a story (I take the liberty to invent a narrative. See Grüne-Yanoff and Schweinzer (2008) for the importance of constructing stories and narratives in interpreting decision and game theoretical models). Let us imagine we are on a quest to seek the Holy Grail and as we walk down a road surrounded by a dark forest, we find ourselves at a fork in the road and have a choice between going left or going right. If we have no prior information, then we cannot necessarily make a good decision between going left or going right. Thus, for all intents and purposes, our choice to go left or go right is random. However, let us say that we do have prior information. Let us say we met a mysterious knight, whose name is Knight Frost, at a tavern in the last town we visited. Over a pint of ale, the knight recounted to us a poem that we now believe is a clue about which path we should select. Based on this information, we decide to take the road less traveled, which is on the left. But since it is the road less traveled, it is a rough, underdeveloped road and results in a great effort to traverse it. Since we expect the payoff to be great, we take it.

The story illustrates that when we deviate from our natural tendency to reduce our probable average rate of work, we may do so only if the expected payoff is great and we may only have such an expectation if we have the requisite prior information and the proper risk attitude. The problem is that we know that researchers do have prior information when they make decisions about which choice they are going to make when they initiate a search. Since we know that, we are left with the notion that if researchers, in aggregate, more often choose one path over another, they do so either because the cost is low and the payoff is at least moderate or the cost is high and the payoff is high enough to justify it.

1.4 Problem Statement

The preferences, utilities, risk attitudes, and prior information held by information seekers and users all play a role in their decisions among a set of relevant alternatives. Unfortunately, it is more difficult to win the hearts and minds of these users and alter their preferences and needs than it is to meet those users where they currently operate. Librarians often contend that users do not take full advantage of their services or collections. The predominant response to this has been to teach users certain skill sets or ways of thinking critically about information and its sources. This response is most representative in the drive to promote and teach *information literacy* skills (ACRL, 2000).

While information literacy skills may be important, and while possessing the skills may encourage the critical evaluation of sources and help ensure the use of good, quality information, this does not entail the use of the library for those sources, and it does not necessarily, given preferences and so forth, encourage that use. As more scholarship and data migrates or is born digital, if it is freely accessible at zero marginal cost to the information user, then such good quality information sources may exist outside the collections of the library and users may use and go to those other sources, for good or ill, despite their skill sets.

If good quality information exists outside the bounds of the academic library and can be discovered with the use of non-library discovery services, then a problem exists if librarians define themselves as primarily about the tools and collections they provide. This may be especially problematic if the tools and collections academic librarians provide become less used in comparison to the tools others provide to locate information outside the bounds of the academic library. The consequences are strategic and can be illustrated with the following set of three chained *modus ponens* (MP), or inferences, which form this dissertation's problem statement.

MP 1:

> P_1: If academic libraries are places where, historically, scholars have acquired most of their scholarly documentation, then academic libraries are places that have had a monopoly on scholarly documentation (Sapp & Gilmour, 2002/3;

Hamlin, 1981; Shiflett, 1981; Wiegand, 1990).

P_2: Scholars can now acquire scholarly documentation from a number of other places (Tenopir, King, Spencer, & Wu, 2009).

C: Therefore, academic libraries no longer have a monopoly on providing scholarly documentation.

MP 2:

P_1: If academic libraries no longer have a monopoly on providing scholarly documentation, then academic libraries are in competition with other places (or other entities) that scholars use to acquire scholarly documentation (Sennyey, Ross, & Mills, 2009).

P_2: Scholars are using these other places (or other entities) as or more frequently than academic libraries for acquiring scholarly documentation (Niu & Hemminger, 2012; Schonfeld & Housewright, 2010).

C: Therefore, these other places (or other entities) are possibly out-competing academic libraries as providers of scholarly documentation.

MP 3:

P_1: If other places are possibly out-competing academic libraries as providers of scholarly documentation, then these other places have dominating strategy profiles.

P_2: Successful competition is largely determined by the choice of a dominating strategy profile (Binmore, 2007; Dixit & Skeath, 2004).

C: Therefore, academic libraries are possibly competing with dominated strategy profiles.

The academic library has played a central role in the life of the researcher and scholar for most of the 20th century, but today these researchers and scholars have other options available to them, and these options provide competing, that is, similar, services and sources of information. The first two *modus ponens* illustrate this and the conclusion expressed in **MP 3**, that academic libraries are possibly competing with dominated strategy profiles or profiles that are strategically weaker than the profiles employed by competing agencies or entities, explains the actions made by researchers and scholars who actively choose these other services and sources of information instead of those provided by librarians. If academic libraries must compete, or are competing, then it must be shown that they are doing so and how.

1.5 Research Questions

Based on the availability of non-library discovery services, such as Google Scholar, the availability of freely accessible content, such as open access journal articles, as well as an aggregate preference for least effort and other decision-making factors such as subjective utility, this study asks and addresses the following two research questions:

1) *Is the current state of affairs, at the network level, such that non-library electronic discovery services marginalize academic libraries?*

The first research question has a strategic dimension, which is highlighted in the following forms:

a) ${}^{1}R_{1}$: Using a third party discovery service to retrieve open access or freely available content is a relevant alternative to using the library's services to retrieve subscribed content.
b) ${}^{1}R_{2}$: The relevant alternative is a competitive alternative.

The second research question, by acknowledging the existence of open access content, grants viability to the strategic dimension of the first research question:

2) *Does open access content, in conjunction with non-library electronic discovery services, marginalize academic libraries?*

The research questions are answered by deriving two operational questions, where the first operational question addresses research question one and the second operational question addresses research question two.

i. What is the probability that any given researcher can use Google Scholar to retrieve a full text document without the benefit of an academic library's proxy or similar service?
ii. What bibliometric or publishing characteristics are driving full text access to

journal articles that users collect?

1.6 Limitations and Delimitations

Open access has a specific definition, but for the purposes of this paper, I use a much broader interpretation of the term to signify anything that is freely available in full text format via an alternative discovery network such as Google Scholar. I apply this weaker definition because of the difficulty in determining the open access status of each full text document found in this study when that document may come from a variety of sources, such as publisher web sites or personal, academic web sites. This also means that I have no preference for the quality of the full text document and thus whether that document is a pre-print, post-print, a copy of a published article, or a word processing document.

Furthermore, the nature of the study is largely limited to the scientific disciplines, and this is due to the nature of the community of users on CiteULike (the population), which itself is inferred primarily by the references sampled and used in this study. Often studies in scholarly communication focus on specific fields of study since each field practices in unique ways with, for example, various emphases placed on certain forms of communication (Wainer, Oliveira, & Anido, 2011), speed of communication (Nicolaisen, 2007), or half-life or obsolecence (Line, 1970) of the products of communication. This study does not attempt to control for this aspect and instead randomly samples from a community most likely composed of researchers from disciplines that lean towards the life, computer, and information sciences.

Although we draw upon citation analysis and bibliometric methods to guide this study, because the object of study, the bibliographic reference, is the same, this study differs because of the context of that reference. Normally the reference is cited by an author, but in this study the reference is collected by a potential reader. Although at least one study has been conducted on the social collecting of bibliographic references and what this activity means with respect to scholarly communication (see Borrego & Fry, 2012), and although the *altmetrics* movement argues for evaluating additional sources of influence (see Priem & Hemminger, 2010), there does not appear to be any strong theoretical study that compares this activity to the citing of a reference. It is hoped that this research will offer some theoretical leads about the behavior and meaning involved in collecting bibliographic records, including whether the actions involved in collecting a bibliographic reference are theoretically comparable to the actions involved in citing a bibliographic reference (Narin & Moll, 1977).

Lastly, this study makes assumptions about the nature of library collections in general. We might posit that as long as an academic library can provide access to a bibliographic record, and not necessarily to the full text document the record refers to, then the library has added that item to its collection, or provided access to it in the same way that print indexes or union catalogs provide listings and possible awareness of publications even if the library has not acquired the specific publications. However, imagine a world where all journals are open access (e.g., Pomerantz, 2006). In such a world, the academic library's core function changes drastically from providing a collection to providing a link to an item where the link is provided by a third-party and the item is stored by a third-party. In this sense, the nature of a library's collection seems to somehow depend on the ratio between directing a user to an access point and both directing and storing an item.

1.7 Definition of Terms

The following definitions provide some grounding for how they are understood more generally here:

- *Academic libraries* --- Academic libraries are libraries in colleges and universities. Historically, academic libraries, especially at research institutions, have had two main purposes: "completeness and control" (Smith, 1990, p. 9). This means that such libraries, at least ideally, seek to acquire, organize, and provide access to all of the scholarly record.
- *Bibliometrics* --- Bibliometrics is a methodology that includes a variety of tools used to examine patterns in scholarly communication. Such methods study particular aspects of the scholarly record, such as the journal article, as well as its use, such as the citation. As well as a staple of library and information science research, Borgman and Furner (2002) write that it is an accepted tool in the sociology of science. Specific bibliometric methods include citation analysis, which is of special interest to those who wish to develop a general theory (Cronin, 1984) or a social theory (Nicolaisen, 2003) of citing as a measure of some kind of influence.
- *CiteULike* --- CiteULike (http://www.citeulike.org/) is a specialized social bookmarking service particularly tailored to meet the document management needs of researchers and scholars (Hull, Pettifer, & Kell, 2008; tbogers, 2009).

Unlike other social bookmarking services that encourage users to capture and tag a link to any web page, CiteULike's focus is scholarly bibliographic references. Essentially, it "is a Web-based tool to help scientists, researchers and academics store, organise, share and discover links to academic research papers," and it has been available since November 2004 (Emamy & Cameron, 2nd para., 2007). Users maintain digital libraries of their collected references, attach memorable tags to these references, and upload articles for later access. Personal libraries are public by default, although users can make their bibliographic references private, and users may form groups based on research interests or projects. These libraries are also indexed by search engines, such as Google and Google Scholar.

- *Google Scholar* --- Google Scholar is a bibliographic database owned by the Google search company. As a bibliographic database, it is similar to others that include Elsevier's Scopus and Thompson Reuter's Web of Knowledge, the latter having origins in work done by Eugene Garfield (1955). Google Scholar's strengths and weaknesses are debated, but research suggests that its ability to retrieve links to a wide range of scholarly communication sources is as strong as its subscription counterparts, such as Scopus and Web of Knowledge (Howland, Wright, Boughan, & Roberts, 2009; Chen, 2010). Other researchers have found that it can retrieve high numbers of open access materials (Norris, Oppenheim, & Rowland, 2008).
- *Institutional repositories* --- An "institutional repository is a set of services that a university offers to the members of its community for the management and

dissemination of digital materials created by the institution and its community members" (Lynch, 2003, p. 328). Lynch argues that institutional repositories can provide a greater dissemination of both traditional (e.g., journal articles) and new forms of scholarly communication. These repositories are often administered and staffed by an institution's academic library. Furthermore, there is an issue of whether academic institutions should mandate author archiving in institutional repositories (Pinfield, 2005; Tibbo, Clemens, & Hank, 2009).

- *Open access* --- Open access refers to "literature [that] is digital, online, free of charge, and free of most copyright and licensing restrictions. What makes it possible is the Internet and the consent of the author or the copyright holder" (Suber, 1st para, 2004). Free of charge refers to the cost to the end user to retrieve the open access document, but there may be costs paid by the author when publishing open access content. Open access publishing is referred to as G*old* Open Access. Open access archiving is referred to as G*reen* Open Access. Green OA does not generally entail any costs to the author and generally refers to archived content (c.f., Harnad et al., 2008), such as pre- or post-prints.
- *Scholarly communication* --- Scholarly communication is a term used to describe the processes scholars, researchers, and professors use to communicate and incorporate the work they do in their positions. It is tied to a process involving four essential actors: researchers/writers, publishers, collectors, and readers. Depending on the various circumstances, additional actors can include granting agencies. Borgman and Furner (2002) also include "peer reviewers, editors,

indexers, information seekers, and readers" (p. 6).

2 LITERATURE REVIEW

2.1 Introduction

This study examines the impact that two states of affairs, non-library resource discovery services and freely available content, have on academic libraries. I propose that since these two states of affairs raise the possibility of bypassing an academic library's services and collections, they have the potential to marginalize academic libraries in at least two of their core functions: collection development and user services. The issue highlights the nature of what it means for a library to collect and to disseminate its collection.

In order to study the issue, the first section of this chapter explores some historical aspects of the academic library and seeks to explain how perspectives of the academic library have shifted in the last century and a half. Since the common perception of libraries is very much intertwined with the collections librarians store and manage, I place particular emphasis on the significance of the library collections. Since developing and managing collections is an activity conducted by librarians, I place additional emphasis on librarianship as a profession.

The second section reviews Google Scholar and outlines how it has become a viable scholarly information discovery service. This involves reviewing the literature that has examined Google Scholar's ability to locate and retrieve scholarly information. Google Scholar would only be partially interesting if it was not possible to retrieve open access content using it. Therefore I follow with a discussion of issues in scholarly communication and publishing. Specifically, rising journal costs and the move to digital formats have spawned and fostered a movement to make scholarly communication freely accessible. As such, I explore the open access movement and briefly discuss its characteristics and outline why researchers and librarians consider it to be important.

The third section outlines the theoretical and methodological dimensions of this study. Since scholarly information behavior is simply another way to refer to the choices scholars make in searching and using scholarly information and since these choices influence the choices made by others, such as librarians, I begin this section with a discussion of the main theoretical framework used in this dissertation: decision and game theory based on bounded rationality assumptions. The theoretical framework is explored using methods developed in bibliometrics and so I follow this with an overview of bibliometrics and citation analysis as tools to study scholarly communication. I report on recent findings that include the use of the web-based and social computing technologies that help expand our understanding of how scholars communicate. Such data sources are citation-based, in that they seek to connect the totality of scholarly communication, under various quality constraints, by capturing the references scholars use to cite to other works. Other data sources, such as those found in social bookmarking web sites, are both broader, in that they allow us to examine what scholars collect and not just cite, and singular, in that what is collected is observable. This latter development is possible because of advances in web technologies and as such it is based on the adoption of certain technologies that impact scholarly information behavior. Therefore, in the fourth section, I describe the use of social computing tools that offer new insights in scholarly communication.

2.2 The Purpose of the Academic Library

The definition of the academic library is an evolving and contested issue, and this is largely due to two issues: the role the library has played in the development of the modern university, as well as the role of the librarian in that setting, and the development of librarianship as a profession (Hamlin, 1981; Shiflett, 1981). In this section, I briefly describe some of the historical discussions related to the development of the academic library. This entails an explanation of the meaning of the library's collections, as it has been understood and discussed in the last century, and an explanation of librarianship as a profession. These two factors, collections and the profession, contribute the most substantial practical and theoretical considerations to the issue involved with defining the academic library, and this is largely because the development and the meaning of the academic library's collection is closely intertwined with the development and meaning of librarianship as a profession. That is, *sine qua non*, one does not make sense without the other.

2.2.1 The Academic Library and its Collections

Historically, the rise of the academic library in the United States began in the 19th century. Wiegand (1990) argues that during this time an ideology of reading, of how and what to read, although often associated with early American public libraries (Ross, 2009), fostered the shape of scholarly communication and academic life, in general. Through much of the 19th century, college curricula remained fairly static. It demanded that students engage, memorize, and translate Greek and Latin works. For those managing libraries at the time, generally faculty and not librarians, this meant that collections need only support a limited canon (Hamlin, 1981). According to Wiegand, this changed after two events: when Charles Darwin published the *Origin of Species* in 1859 and when the United States passed the 1862 Morrill Act, which set aside lands for colleges to study agriculture and the mechanical arts. In addition to the research library movement (Shiflett, 1981), these two events upset previous pedagogy and curricula, challenged established assumptions about the purpose of the academy, and contributed to a "culture [which] consisted of experts whose job it was to find new truths to replace the old authority patterns" (Wiegand, 1990, p. 74). Hence, the revolution involved developing and exploring new sources of data and methodologies, which lead to an emphasis on the creation of new knowledge. While it might be argued that the events described by Wiegand were insufficient for such change, that other events contributed to the "quest for fact-gathering" (p. 75), and that the changes were in play long before by events such as the Copernican revolution and the invention of the printing press (Eisenstein, 1983), it does seem true that the new environment, the new conditions, and the change in world-views accelerated "demands for more data [which] led to the acquisition of major research collections" (Wiegand, 1990, p. 75). These demands led to new journals and eventually to new responsibilities for librarians, such as collection development. Wiegand notes that "often that data is found in research libraries" (p. 81) in the form of its collections.

For academic libraries, the focus on developing comprehensive collections continued through most of the 20th century. However, this purpose of academic libraries rests on fundamental questions about what a collection is, what form it exists in, and how the items in the collection are transmitted, stored, and retrieved. In 1978 F. W. Lancaster published the controversial and discussion-provoking work *Toward Paperless Information Systems*. Lancaster predicted that by the end of the 20th century, automation and other technological developments would lead to a society where the primary mode of communication, and especially scholarly communication, would be electronic. Lancaster's argument, in part, arose from certain trends in academic libraries and scholarly publishing at the time. He notes that in the early 1970s, academic libraries were able to keep pace with the amount of published scholarship, at least in terms of titles if not volumes, but due to rising costs in serials and book titles, as well as personnel costs, which included the labor intensive activity of collection management, the trend would not be sustainable through the end of the 20th century. Lancaster writes that "in the period 1965-1972, total expenditures of the great research libraries increased some 103% and materials and binding expenditures by some 78%. Yet the collections grew in size by only 37% in this period" (p. 94).

As a result of the creation of the web in the early 1990s and the rise of published scholarship via this medium in the intervening years, Lancaster's prediction about a paperless society has turned out to be mostly true, in a complicated fashion, and the result leads to the implication of what it means to have a collection that is digital in form. At the heart of the issue is the idea of a paperless society and the ubiquitous availability of personal search, retrieval, and storage devices, as envisioned by Vannevar Bush (1945) and J. C. R. Licklider (1965). The implication concerns whether such a society would render the academic library obsolete given that one of its core functions has been to develop and maintain comprehensive print collections.

The dawn of library automation in the 1930s (Parker, 1936; Kilgour, 1939; Black, 2007) launched an era of predictions about the future of academic libraries. After Licklider and others warned librarians about the potential implications of a paperless society, and what that meant for libraries where a core service was building a print collection, Sapp and Gilmour (2002; 2003) note that the literature written by librarians and library and information scientists began to shift away from a focus on collections and towards a focus on the users. Instead of a future where "Libraries could not and should not expect to retain a monopoly over information" (Sapp & Gilmour, 2002, "The Next Decade in Academic Librarianship," para. 3), librarians should expect to adjust to a future where information is decentralized and where other information agencies, including for-profit ones, have much more direct control over the dissemination of content to end users. Sapp and Gilmour (2002) write that in 1985, Allen B. Veanor, a library consultant commissioned by the Association of College and Research Libraries (ACRL), argued that "The breakup of the academic library's monopoly on information inevitably would result in competition from external, non-academic entities. This would cause an increasing number of information resources to be marketed directly to the user" (para. 5).

Arguments about the competitive role of the academic library have been made by others and more recently. Sennyey, Ross, and Mills (2009) describe changes for the academic library as it moves into a landscape dominated by digitized and digital collections. They note that digital and digitized content, and especially open access content, "creates a growing corpora that is accessible outside of the aegis of the library"

(p. 254) and this puts the academic library into a competitive relationship with others in the scholarly communication system, such as publishers. They note that the competition is for patrons.

The rise of digital content and the orientation towards the library user have made an impact on what it means to collect. Harloe and Budd (1994) argue that content, and not packaging, should drive collection management. They make the case that the needs of the community are paramount, and quoting Sheila Dowd (1990), write that,

> 'Bits and bytes of information are important only if the mind can link them with other pieces of information to build the orderly patterns that are fabric of knowledge. Hence the mission of the library is more properly identified as the provision of access to organized information, *for the fostering of knowledge*' [emphasis added] (p. 87).

Despite the cognitive and epistemological emphasis on what a collection means by authors such as Harloe and Budd (1994), others in the field continued to emphasize the importance of the physical collection. Carrigan (1995) argues that the primary purpose of the library is to offer certain benefits to its users and the greatest of these benefits is its collection. He writes that "Libraries have multiple functions but all functions presume ultimate use of libraries' collections" (p. 100). This statement highlights perhaps the most important premise traditionally held by academic librarians---and as such, it is a crucial premise in an argument that collections are one of the core services a library provides. Carrigan argues that a collection should be evaluated by its return on investment, but a return on investment should not be measured by use, but by whether the collection fits the needs of the community.

Though the academic library has a contested definition and purpose, what is clear is that a balancing act exists between the role of the library in developing or managing collections and the role of the library in the life of the user. Akeroyd (2001) argues that "It is all about becoming more user centered and less collection focused or function dominated" (p. 82). As an example, the library at the University of North Carolina at Chapel-Hill (UNC) has undergone substantial changes in the last seven years, and these changes visit the heart of what defines an academic library. Michalak (2012) describes some of the most significant of these changes, which include a description of the UNC library as "outward facing" (p. 412), meaning that not only have collections become less the purview of the library, as content has become digitized and decentralized, but so has service. This means that librarians now spend more time going to the academic library user and not waiting for the academic user to come to the library. Michalak finds that service follows the collection, and as the collection has become digitized and decentralized, so has the "service dynamic" (p. 413). Others have observed that the academic library is becoming more of a learning organization (Senge, 1990), and this not only has had an effect on the services offered but also on the organizational structure of the library, which is grounded in "information sharing, team-based structure, empowered employees, decentralized decision making and participative strategy" (Moran, 2001, p. 108).

2.2.2 Librarianship as Profession

Michalak (2012) and Moran's (2001) observations underlie the role of the librarian in the academic library. While automation and digitization have had a substantial impact on what it means to collect and what the nature of a collection is, this impact has not only influenced the academic library's definition but also what it means to be a librarian. When Ralph H. Parker implemented the first library automation project in 1936, the goal was to pursue "a new day of no mistakes, no nervous strain, and much less manual labor for the library worker" (1936, p. 905). Parker's motivation was to create a better working environment for the librarian, one that had a stronger intellectual base than what the mundane tasks required from librarians at the time.

Despite such motivation, librarians have faced considerable obstacles in establishing themselves as a professional class. Part of the issue has been blamed on society's biases towards the feminization of the work. Mitchell (2007) highlights the early discrepancy between the pay given to women who worked in libraries and the pay given to men, and that much of this discrepancy was a result of the clerical work administered to women in the 19th century (and also because of the historical bias against women, such that they are given clerical work instead of intellectual work). In a piece on the status of and conditions for paraprofessionals in American academic libraries, Oberg, Mentges, McDermott, and Harusadangkul (1992) reflect on the lack of self-esteem among academic librarians and their inability to define themselves and their paraprofessional counterparts.

While lack of esteem and respect among academic librarians might be partially blamed on the societal biases placed on professions where there is a dominance of female workers, and also by the nature of the work as service work, Carpenter (1996) adds a different argument. He proposes that librarians have received less stature than faculty because their work has primarily been about the dissemination of knowledge and not the creation of knowledge. Carpenter's view adds to Wiegand's (1990) historical observation about academic libraries, noted earlier, about the change in the purpose of the university during the 19th century, from institutions with a primary mission to disseminate classical knowledge to institutions with a primary mission to create new knowledge: "the more 'pure,' the more highly esteemed" (Carpenter, p. 87). In essence, knowledge creation became the primary virtue of the academy, at the cost of knowledge dissemination.

The decentralization of digital collections and their accessibility outside the *aegis* of the library, the importance of the content of the collection rather than the format, the reaching out to the user rather than passively waiting for them, and the desire to professionalize librarianship imply that the competition for the patron does not lie in a competition over collections. Rather, it implies that the competitive strength of the academic library lies with librarians and their ability to serve their communities in a way that makes strategic sense given their mission. Plutchak (2012) describes the strategic necessity of developing skills that best serve librarians' communities. He also argues that the tendency to personify the library, such that the library performs actions instead of the librarian, only serves to reduce the importance of the librarian's role. It will be important to discern, assuming librarians maintain a focus on knowledge dissemination rather than creation, is whether dissemination becomes a more important cultural value in the age of

Google and in an academic system that involves the decentralized storage of content across many types of entities, such as open access journals.

2.3 Alternate Discovery Services and the Decentralization of Collections

In this section, I describe Google Scholar, which is used increasingly by researchers to search for relevant information. In the second part of this section, I discuss the open access movement and outline some of the reasons why it has become an important topic for both librarians and researchers. I use these sections to provide justification for the later analysis of the overall issue, which is that it is possible and rational to use Google Scholar to acquire open access content and that this means that it is possible to not use the academic library to acquire content from its collections.

2.3.1 Google Scholar: Alternate Discovery Services

Google Scholar (GS) has become an important bibliographic database and citation index as it has become more capable of indexing a comprehensive amount of scholarly documentation. Unlike Scopus or Web of Science, it is freely available to any user with an Internet connection. Studies show that GS is perceived to be useful to end users (Cothran, 2011), is becoming a growing presence on academic library web sites (Neuhaus, Neuhaus, & Asher, 2008), and is becoming a preferred choice among academic library users, especially those in the sciences and the social sciences if not necessarily those in the humanities (Herrera, 2011).

According to a study conducted by Baldwin (2009), GS "indexes publisher web sites, PubMed Central (PubMed), institutional repositories, preprint archives, etc. It also locates full text results from research groups posting articles online for their own use and failing to make access proprietary" (pp. 3-4). Baldwin's study suggests variability in sources used to retrieve full-text documents depending on the type of article and subject matter being searched. For example, in a comparison between searches for mechanical and chemical engineering, nine percent of mechanical engineering full-text articles were sourced from PubMed whereas 48% of chemical engineering articles searched originated from PubMed. Institutional repositories provided a nearly even balance between the two searches (10% and 13%, respectively), and seven percent of the found mechanical engineering articles were sourced from publishers' open access sites compared to 27% of the found chemical engineering articles (p. 6). In a study by Meho and Yang (2007), as cited by Harzing and Wal (2008), there is a small overlap between the subscription databases Web of Science and Scopus with GS: "The overlap between WoS and Scopus was 58.2%. The overlap between GS and the union of WoS and Scopus was only 38.8%" (p. 3).

In a study to evaluate "the breadth and scope of available content" on GS, Howland et al. (2009) recruited seven subject librarians to conduct subject directed queries covering topics in the sciences, the social sciences, and the humanities and compared these results using subscription databases at Brigham Young University. The results revealed that the "mean scholariness [*sic*] score of citations found in Google Scholar was 17.6 percent higher than the score for citations found only in licensed databases" (p. 231). Additionally, the researchers discovered that the results for all three subject areas were balanced and that there was no statistical difference in their "scholariness score."

Additionally, Howland et al. discovered that, when conducting specific title searches, "Google Scholar actually contained 76 percent of all the citations found in the library databases, while the library databases contained only 47 percent of the citations found in Google Scholar" (p. 231). The library databases used in this study included SportDiscus, Medline, Applied Sciences and Technology Abstracts, Business Source Premier, PsychINFO, Linguistics and Language Behavior Abstracts, and JSTOR. While such studies suggest the strengths of GS, it should be noted that coverage of all disciplines is not universal. Kirkwood and Kirkwood (2011) find mixed results in GS's coverage of historical scholarship, and institutional repositories using the Dublin Core Metadata Element Set can be overlooked by GS given certain deficiencies in the ability of Dublin Core to appropriately describe scholarly content (Arlitsch & O'Brien, 2012).

In an interesting and perhaps, within its very limited framework, successful attempt to measure recall and precision in GS, within the scope of the subject area searched ("later-life migration"), Walters (2009) found that "GS performs better than many subscription databases" (p. 16). In this study, involving a comparison of GS and eleven subscription databases, relevance was defined as an assessment of "subject matter, importance of findings, innovativeness of methods or approach, number of other studies published on the topic, accessibility of content (readability), and accessibility of the document itself (availability to students and scholars)" (p. 7). One hundred and fifty five papers were selected for the recall and precision study. GS placed fourth in both recall and precision when evaluating the first ten hits and moved to first place after 75 result hits. For the most part, the differences between first, second, third, and fourth place were trivial.

2.3.2 Open Access: The Decentralization of Collections

For the last thirty to forty years, journal prices have increased at a rate that has been difficult for libraries to match. The end result is a situation librarians refer to as the "serials crisis" (Greco, Wharton, Estelami, & Jones, 2006). Some have argued or pointed out that part of the reason for the increased cost in journal prices is due to the costs involved in publishing both print and online formats (Kling & Callahan, 2003; Fidczuk, Beebe, & Wallas, 2007). Others have argued that copyright law creates a monopoly that allows publishers to charge exorbitant fees (Bergstrom & Bergstrom, 2006). While there are certainly other causes, the end result is a system that many believe is unsustainable.

Although academic libraries command a seemingly large budget for the acquisition of materials, the average annual price for serials has increased at a much faster rate than library acquisition budgets. For the 2010-2011 year, the Association of Research Libraries (ARL) reports that "total library expenditures of all 126 member libraries...was slightly more than $4.6 billion" (ARL Statistics: 2010-2011, 2012, p. 5). For member university libraries, 42.80% of those expenditures was spent on library materials with a median cost of $7,451,090 for the 114 reporting libraries on current serials (p. 44). Compare this to the 2004-2005 year, total library expenditures for the then 123 ARL members neared $3.6 billion, representing a 27.78% total increase or an average of 3.97% per year for year to date (ARL Statistics: 2004-2005, 2006). For member university libraries, 40% of those expenditures was spent on library materials and the median cost for serial expenditures for 112 of the 2004-2005 university libraries was $5,904,464 (p. 42). This represents an average annual increase of 3.74%, slightly less than total library expenditures. Despite this, the average annual serials price increase for 2012 journals is 6% (Bosch & Henderson, 2012), but in the past, annual serial price increases have been far higher. McGuigan and Russell (2008) report that from 1986 to 2005, the average annual price increase for all serials was 7.6%. Budd (2002) finds that the average serial price increase for social science titles was 9.4% during the 1990s. Romero (2008) reports that serial prices in communication studies increased by 223% from 1994 through 2004 and by 1,780% for law studies.

Proponents of open access (OA), as a publishing model, have argued that it can help alleviate the burden on academic libraries' serials and acquisitions budgets (Corrado, 2005; Albert, 2006). The ARL statistics highlight how this may yet be the case, and it may also depend on what type of OA model is pursued. Open access exists in two broad forms: Gold OA and Green OA. Lewis (2012) describes the types of Gold OA models. "Direct Gold OA" pertains to journals that publish articles that are freely accessible to readers at the time of publication. Journals that provide access to articles after an embargo period are considered Delay Gold OA journals. Hybrid Gold OA journals give authors an option to pay a submission or publication fee. When authors pay this fee, their articles will be immediately accessible to readers even in journal issues that have articles that are not OA because other authors did not pay a fee.

Green OA, on the other hand, "sits alongside the subscription journal system and does not attempt to replace it" (Lewis, 2012, p. 494). This model is primarily about self-archiving the publication. Authors who take advantage of Green OA have several options for self-archiving. They may deposit a copy of the article's pre-print or post-print version either on their personal web site or in an institutional or subject repository. Pre-prints are versions of the article that have yet to be peer-reviewed and post-prints are versions of the article that have been peer-reviewed. Chan (2004) distinguishes between Gold and Green OA as open access publishing (OAP) and open access archiving (OAA), respectively. Both OAP and OAA models are original definitions in the Budapest Open Access Initiative, which was released in February 2002 and provides the core definition of open access (Bailey, 2007). Other OA characteristics noted by Bailey include content that is freely available, is online, and has minimal restrictions for re-use. The re-use factor relates to copyright, which is often held by the author(s) of an OA work and may be assigned a Creative Commons license.

Open access research largely focuses on three areas: the benefits to libraries in the form of journal cost-saving, the benefits to the public and to scholars in the form of increased access, and the influence of open access in terms of citation counts or number of downloads, which is a possible indicator of readership. While the first two types of research focus on the implications of open access for libraries and readers, those implications are often one-sided. That is, it is assumed that the benefits outweigh any costs, where the costs might be the marginalization of academic libraries, in terms of the decentralization of content storage, or some other unnamed implication. Drott (2006), for example, illustrates that "the emergence of the discussion of open access as a viable alternative to traditional publishing rests on developments in three main areas: economics, technology, and social justice (p. 81). Thus, while OA's impact on libraries' budgets is often a major component of the discussion, the impact on the use of the library's collection is not.

Research that focuses on measuring OA's influence by comparing downloads and citations between open-access and subscription-only articles or journals includes as its audience other researchers with interest, for various reasons, in such measures when deciding to publish in open access or subscription-based journals. Generally, this research suggests that open access articles have increased download rates, but there is no agreement that open access articles have a citation advantage---an increased likelihood of citability or an increased citation count. For instance, in a randomized controlled trial involving journals published by the American Physiological Society, Davis, Lewenstein, Simon, Booth, and Connolly (2008) found open access articles led to substantially increased downloads over subscription-only articles with 89% more full text, open access downloads, $r^2 = .42$, $F(31, 1350) = 37.5$, $p < 0.001$, 95% *CI* [34 to 53] (p. 4-5). However, they found that, after one year, the access level had little to do with citability: 63% of the subscription-only articles were cited and 59% of the open access articles were cited. This finding is in direct conflict with Eysenbach (2006), whose study of a single journal (*PNAS: Proceedings of the National Academy of Sciences*) found that after a mean of 206 days plus six months after publication, subscription-only articles were less often cited than open access articles. Specifically, 51% of the subscription-only articles were cited in contrast to 63.2% of the open access articles (relative risk = 1.3, 95% *CI* [1.1 to 1.6]). Eysenbach also found that open access articles saw a higher citation count as early as four months after publication. Between six and ten months after publication, open access articles versus subscription-only articles, respectively, received average counts of 6.4 [*SD* = 10.4] and 4.5 [*SD* = 4.9]; $z = 4.058$; $p < 0.001$.

However, Gargouri et al. (2010) found an open access citation advantage primarily for higher quality open access articles, which saw nearly an eight fold odds increase in citation counts, *DV* = 1-4 cites (low) vs. 20+ cites (high), *OR* = 7.953. The study examined subscription-only articles, mandated institutional repository open access articles, and self-selected open access articles. It specifically compared subscription-only articles against self-selected open access articles, subscription-only articles against mandated institutional open access articles, and self-selected open access articles against mandated institutional repository open access articles. Gargouri et al. concluded that high quality articles see many more citations if the articles are open access. They ruled out a self-selection bias, which has been put forth as the argument that if open access sees a citation advantage, it is caused by authors self-selecting as open access their best work. Instead, they infer that there is a "quality advantage" due to "user self-selection" (Discussion section, para. 5) and not author self-selection.

Whether there exists a download or a citation advantage, these studies demonstrate OA's influence on the research front. However, the growing number of OA journals mean that academic libraries do not always provide records to open access journals in their catalogs. Additionally, the main bibliographic indexes, including Web of Science, EBSCO Academic Search Complete, ProQuest Research Library, Biological Abstracts, and others do not always list open access journals and those journals that are listed generally have privileged characteristics, such as high impact factors and high publication output per year; they may also be U.S. based and charge authors fees to publish (Collins & Walters, 2010; Walters & Linvill, 2011a; Walters & Linvill, 2011b). This perhaps entails that much OA published content is left to be discovered by less discriminate services, such as GS. Despite the disagreement among the findings and the uncertain accessibility of OA content in library supplied databases, these studies suggest that OA content has an increasingly broader reach than articles that exist behind a pay wall and that this is in large part because services such as GS are good at locating OA content.

2.4 Theoretical and Methodological Bases for the Study

If researchers use non-library services, such as GS, to acquire documentation, such as OA content that does not necessarily have to be collected by libraries, and if such actions have implications for the academic library, then it does so and is possible for several reasons. First, decisions about where to begin a literature search, such as the academic library's web site or GS, represent decisions made by people. As such, I begin this section with a discussion of decision and game theory as well as the notions of rationality upon which these theories hinge. The purpose of this section is to show that it can be rational not to use an academic library's services and collections. By rational, I mean that the payoff for the scholar who uses non-library services to retrieve non-library documents is sufficient enough to warrant the continued use of those services. If academic librarians are to respond to these actions, the justified rationality of the searcher will have to be taken into consideration.

Since this study gathers data from a social computing web site where users of the web site collect and store bibliographic references, and since these bibliographic references are analyzed using bibliometric data collected from GS, I follow with theoretical discussions of bibliometrics, social computing, and what it means to collect bibliographic references. Essentially, while the act of citing a scholarly document with a bibliographic reference has been a primary object of study in information science for the last fifty years (Narin & Moll, 1977), the act of collecting and saving bibliographic references to scholarly documentation on a social computing web site represents an activity that is just beginning to be explored. However, citation theory may be used to build a framework outlining what it means to collect a bibliographic reference, in terms of whether the social activity involved with collecting a bibliographic reference suggests something meaningful about the document that is referenced in an analogous way that citing suggests something meaningful about the relationship between a citing and a cited document. Furthermore, the ability to collect these references on the social computing web sites built for such purposes contributes a necessary theoretical part of this study. This ability is only possible and is only acted on because of certain technological affordances offered by these social computing web sites.

2.4.1 Decision, Game Theory, and Bounded Rationality

Decision theory describes those "situations where each person can choose without concern for reaction or response from others" (Dixit & Skeath, 2004, p. 18). Game theory describes those situations where decisions by a player interact with decisions made by other players. It has been used to explain topics in economics, political science, sociology, and philosophy (De Bruin, 2005). Dixit and Skeath (2004) outline several components of strategic games. Players have strategies, where these strategies are simply the relevant "choices available to them" (p. 27). The outcomes of a game are described as payoffs, and these are usually assigned some numerical score (such as the number of dollars awarded for some outcome). Additionally, the players are thought or assumed to be rational in that they seek to achieve the highest payoff. Last, all strategic games have solutions which are described in terms of the game's equilibrium. An equilibrium in a game "simply means that each player is using the strategy that is the best response to the strategies of the other players" (p. 33).

In economic game theory, utility represents the value a person places on some thing. The value in attaining the utility is expressed as a payoff. In the Prisoner's Dilemma (PD), utility is expressed in reduced or no years served in prison and the payoff of the game is expressed in how many prison years are served, if any, as a result of playing the game against a competitor. Hausman (2005) refers to utilities as "indices of preference" (p. 36). Mapping preferences to utilities can be arbitrary. Binmore (2008) suggests taking the most important preference and assigning, e.g., a 100 to it on a scale of 1 to 100, and taking the least important preference and assigning a 1 to it. That method serves to define the upper and lower bounds of the utility function (c.f., Kruschke, 2011, pp. 28-29). However, according to Dixit and Skeath (2004), oftentimes these "numbers are only educated guesses" (p. 28).

In the economics and philosophy literature on game theory, there is often a conflation between utilities and preferences, and this conflation can be a matter of contention. The dominant theory in economics is a notion of preference as choice ranking, which is informed by Paul Samuelson's theory of revealed preference (Binmore, 1994; Hausman, 2000). An easy example of revealed preference follows: if I go to a grocery story and see before me a basket of apples and a basket of oranges, and if I choose an apple, all things being equal, then revealed preference says that I prefer apples over oranges.

Amartya Sen (1973; 1977) argues that preferences are more complex than the reduced/simplistic notion of preference as defined above. Sen argues that there are different kinds of preferences. A list may include or take into consideration beliefs, expected advantage, desirability, social norms, moral principles, habits, sympathy, commitment (Hausman, 2005). While Hausman agrees that preferences are more complex than rational choice theory suggests, he argues that there should be one theory of preference that takes into consideration or includes all these.

The value we assign to preferences are mapped to utilities, and what we [should] choose, based on our utility functions, are strategies (Hausman, 2005). There are two valuations of strategies: dominate and dominated, and there are degrees of both, so that a strategy *A* may strictly or weakly dominate strategy *B*; or a strategy *B* may be strictly or weakly dominated by strategy *A*. A dominate strategy, one that is rational, may not always equal a fair strategy or a collectively rational strategy (Binmore, 1994). The strategic game called the Prisoner's Dilemma (PD) illustrates a scenario where the dominant strategy, the best strategic response to someone else's best strategic choice, results in a worse payoff for both players.

Games are classified as either cooperative or non-cooperative and all games have solutions. An equilibrium, Nash or otherwise, is a solution to a game, where the solution is the game's stable state, which means no player is better off switching strategies (Ross, 2010). Both types of games can reach stability. The main difference between cooperative and non-cooperative games is that cooperative games model coalitions (Rosenthal, 2011). In non-cooperative games, players can agree to cooperate, but they cannot trust that other player(s) will.

Game theory is applicable in a descriptive way (Cave, 2005) as it relates to identifying preferences as those preferences exist in odds or in agreement with a payoff. For example, librarians prefer lower subscription rates for their serials although they continue to pay higher costs. If, though, librarians are rational agents and make rational choices, but are still not acquiring the payoff they desire, then by holding their rationality constant, game theory can help explain why librarians are not achieving their desired payoff. In such cases, for example, it may be that they are coerced into playing with dominating strategies.

One problem with traditional game theory is the strict assumption it has about rationality (Budd, 2012). Here rationality often means assuming that players in a game have complete knowledge of their own preferences and are able to perform "flawless calculation[s] of what actions will best serve those" preferences (Dixit & Skeath, 2004, p. 30). Additionally, it also generally means that players will remain consistent about their preferences, such that, if a player has a preference for oranges that is greater than a preference for apples, then the player will always choose an orange and not an apple when presented with both (see Ritzberger (2002) for a discussion on rationality assumptions such as completeness and transitivity).

Consider, for example, the *Ultimatum Game*, where two players, a Proposer and a Responder, must decide how to split a pot of money. In this game, imagine a $20 pot of money from which the Proposer must offer a part to the Responder. Both know that if the Responder rejects the offer, neither receive any payoff. If the Responder accepts the offer, they receive a share based on the proposed split.

The rationality assumption often adhered to by game theorists means that even if the Proposer offers the Responder a $1 in order to keep $19 for himself, the Responder will accept this offer since receiving some money is generally better than receiving no money, given that the Responder has a preference for more money. That is, the Responder is selecting his best strategy given the strategy selected by the Proposer. As such, a $1 and $19 split represent a solution to the game, otherwise referred to as its equilibrium. However, studies show that "the majority of proposers offer 40 to 50% of the total sum, and about half of all responders reject offers below 30%" (Nowak, Page, & Sigmund, p. 1773, 2000). Common explanations for this behavior incorporate notions of fairness, reputation, and retribution even though these represent affective states, social norms, and not rational attitudes.

The same kind of rationality assumption can be applied to the study of scholarly information seeking. Consider that a researcher requires information about topic *X*. Simplifying the strategies available to the researcher, suppose that the researcher has two: one strategy is to use the library's resources to acquire a document about *X* and the other strategy is to use a non-library resource to acquire a document about *X*. The payoff for either strategy is access to a relevant document about *X*. Given that the payoff is the same for both strategies, the difference between the two strategies exists in terms of the costs to the researcher. Essentially, if it will cost the researcher less in terms of time, knowledge, or frustration with the retrieval systems, to use either the library resource or the non-library resource, then the researcher will always use that strategy which will cost him less. Thus, the payoff to the researcher is the value of the relevant information minus the cost in acquiring that information. The researcher, in this game, will want to incur as little cost as possible.

In order to define the game, it is necessary to take measure of competing agencies abilities to offer their services. For example, if choosing GS is a strategic option for the researcher, then GS must provide the necessary tools for it to be a rational option. The same holds true for the library. However, in either case, both players must be successful in returning and providing access to relevant information upon request. That is, the payoff must be the same and only the incurred costs may vary.

How a researcher may choose between these two options may very well depend on what he believes is the best option, given that the researcher may have incomplete or incorrect information about the ability of either the library or the non-library service to succeed in returning and providing access to the relevant information. Psychological game theory (Dufwenberg, 2010) suggests that "belief-dependent motivations," where the game's payoffs "are defined on beliefs (about actions and beliefs), as well as on which actions are chosen" (p. 272), might shed light on the researcher's strategy profile given the prior beliefs the researcher may hold about the strategies available to him. If the researcher often uses the general Google search service, then because it is a service provided by the same brand name, the researcher may carry with him a belief about GS and its ability to satisfy his information needs. Given the success of Google, such a belief may provide an advantage to GS. This advantage may be more pronounced if the researcher has, in the past, developed beliefs about the library's services. If the researcher has ever found the library's services wanting or frustrating to use and navigate (Yadamsuren, Paul, Wang, Wang, & Erdelez, 2008; Kress, Bosque, & Ipri, 2011), given his beliefs about the non-library service, then the library service will be at a disadvantage, at least in the mind of the researcher.

This problem may hinge, for example, on the researcher's perceived cost of either service. Zipf (1949) might argue that the perceived cost will be dependent on not only the amount of work involved in using either service, but the researcher's estimate of the probability of using either service over the long run. For Zipf,

> The most that any individual can do is to estimate what his future problems are likely to be, and then govern his conduct accordingly. In other words, before an individual can minimize his average rate of work-expenditure over time, he must first estimate the probable eventualities of his future, and then select a path of least average rate of work through these.
>
> Yet in so doing the individual is no longer minimizing an average rate of work, but *a probable average rate of work; or he is governed by the principle of the least average rate of probable work*.
>
> For convenience, we shall use the term least effort to describe the preceding least

> average rate of probable work (p. 6).

The least average rate of probable work will be defined by our ability to solve problems and apply search heuristics given our limited computational abilities as human beings. Herbert Simon's (1990) notion of *bounded rationality* flushes out Zipf's *principle of least effort* in the sense that our "computational limitations" in tandem with the characteristics of the systems we use to search result "*not in optimizing techniques, but methods for arriving at satisfactory solutions with modest amounts of computation*," (p. 11) or shall we say, with as little cost or "least average rate of probably work" as possible. For Simon, we do not maximize our utilities; rather, due to our limitations and to our settings, we simply attempt to satisfy our preferences in whatever way it reduces our computational load (incurs less cost to us).

It is, then, necessary to show that what is satisfactory simply refers to what is most probable, or what is believed to be most probable, given the work involved and the setting of the work. When making a decision, a person has at least two options to consider, two ways to act, in order to achieve some outcome. If the person is rational, she will choose the act that will most likely result in the desired outcome. If a person requires a journal article and has before her several paths to acquire the journal article, then she will, we suppose, choose the path that will most probably result in acquiring the article, or the path that she at least believes is the most probable and that requires the least amount of effort. In such a case, we act as Bayesians (Phillips, 1973), starting off, at least implicitly, assigning probabilities based on known prior probabilities or intuitions about unknown prior probabilities, and then updating our beliefs, or our posteriors (Alder &

Roessler, 1968), after we have made a decision and learned what the outcome of that decision is; that is, after acquiring new information.

Not all choices are equal. Any two paths to acquire a document may require different levels of effort and may put before us different barriers or obstructions. As an example, if the probability of using the physical library *P(x)* to acquire a known document is 95% and the probability of using Google Scholar *P(y)* from my home computer to acquire the same document, freely, is 5%, and if I am at home, then it makes rational sense to use GS before I trek down to the library or to use the library's databases if it means having to sign in and navigate through a handful of web pages. If $P(x) = 0.95$ and $P(y) = .05$, but I choose *P(y)* first and assume that I am rational in doing so, then some kind of weight should be applied to *P(y)* to show that it is preferable to *P(x)* in the order of things; that is, $P(y) > P(x)$. Essentially, we should assume that people often have good reasons for the decisions they make and our calculations should reflect that.

In a sense, the purpose of this study is to outline how *P(y)*, using GS, might be rational given that *P(x)*, using the library, often means a sure payoff. It seems that once *P(y)* has reached a certain point, so that it will in the long run return a successful outcome, then *P(x)* can always be a sure thing and yet it will not matter if it means incurring a greater cost to the searcher.

2.4.2 Bibliometrics and Citation Analysis

Broadus (1987) defines bibliometrics as the "'quantitative study of physical published units, or of bibliographic units, or of the surrogates for either'" (p. 376). White and McCain (1989) note that "bibliometrics is to publications as demography is to peoples" (p. 122). If so, then what composes the bibliometric study defines and sets its boundaries. Often, researchers gather bibliometric statistics from citation lists generated by bibliographic databases such as those provided by Thompson Reuter's Institute of Scientific Information (ISI) indexes (e.g., Web of Science). More recently, interest has risen in Elsevier's Scopus and Google's Google Scholar as sources for both bibliometrics and citation analysis (e.g., Noruzi, 2005; Yang & Meho, 2006; Falagas, Pitsouni, Malietziz, & Pappas, 2008; Harzing & Wal, 2008; Howland et al., 2009).

While these data sources differ in scope, they each represent fundamentally the same framework and intent: to capture formal scholarly communication (Wouters, 1998), authenticated or authorized as such in some standard fashion, and to offer some ability to understand the relationships between authors, journals (or other formats), and their communities through their publications.

As methodologies and methods, bibliometrics and citation analysis have been used for a variety of purposes and to develop and test certain theories. They have an object of study, the publication as a whole and its various components including authorship, the byline (Cronin, Shaw, & La Barre, 2003), the reference, and the citation. They have a way of going about what they study---their methods, which may include counting citations, examining author co-citations, and analyzing bibliographic coupling relationships. The motivations for these studies may be practical. For example, McCain and Bobick (1981) used citation analysis to study journal use in an academic library. More recently, Enger (2009) used citation analysis to study core book collections in an academic library in order to further collection development methods.

Often, the motivation for these studies stems out of a theoretical interest in communication, attribution, dissemination, and retrieval as well as an interest in the sociological nature of various scholarly communities. To illustrate, Cronin (1984) provides an in-depth review of citation studies and outlines the need for a citation theory given that scholarly and scientific communication perform important social functions. Bornmann and Hans-Dieter (2008) examine whether citing behavior is a reliable method for examining influence and appropriating credit. Hellqvist (2010) studies what it means to reference a work in the humanities under the explicit assumption that referencing in the humanities is different than referencing in the sciences. Case and Higgins (2000) possess a similar motivation in their study of citation behavior among scholars in the field of communication. Budd (1986) maps the subject area of American literature, and White (2007a; 2007b) describes how bibliometrics can enhance information retrieval systems.

Measurement issues are a concern. Ding and Cronin (2011) note the distinction between popularity and prestige by noting the distinction between being highly cited and being cited by highly cited papers. The attempt is to weigh citations rather than to hold each citation as an equal unit of influence. For apparently similar weighting reasons, Nicolaisen (2002) highlights how some highly cited papers are negatively cited because, for example, they are contested in some way. Meho (2007) points out that most articles are not cited. MacRoberts and MacRoberts (2010) show that some sources of influence, such as data sets, are not cited at all because the norms of citing often devalue these as instants of formal influence. As sources of measurement, Vaughan and Shaw (2003) demonstrate a correlation between citations provided by the standard indexes and citations on the web.

Nicolaisen (2003) writes that "in order to understand, explain, and predict the dynamics of citation networks, we need to penetrate the social worlds of individual authors" (p. 18). This is also true of bibliometrics in general. The problem is, according to Nicolaisen, "not uncomplicated." While penetrating the social worlds of scholars and scientists may be difficult, advances in social computing technologies (O'Reilly, 2005) may offer insight into these social worlds as well as the variety of research traditions that exist around them. Importantly, these insights may be derived from the "empirical grounding" Nicolaisen seeks from a social theory of citing and, by extension, bibliometrics too. Specifically, this empirical starting point may lie at the intersection where social computing and bibliographic reference collecting converge and may exist in supplement to the empirical grounding of more traditional sources such as the Science Citation Index (SCI), as historically outlined by De Bellis (2009). Thus, web-based applications such as CiteULike, BibSonomy and others, where users of these sites collect, store, tag, and share bibliographic references, serve as likely candidates of attention. As Cronin (2001) noted, "the web has challenged, and may revolutionize, many of the assumptions that have underpinned the established scholarly communication system" (p. 3) as well as potentially enabled us "to detect early signs of emerging trends" (p.6).

2.4.3 Social Computing

If the web revolutionizes assumptions about scholarly communication, alerts us to emerging trends, as well as alters our actions, habits, and behaviors, then it does this most effectively through social computing, and in particular, to two important attributes of this phenomenon: place (see also Pomerantz & Marchionini, 2007) and affordance. Dourish (2001) defines affordance with regards to social computing, human-computer interaction, and system design as a "a property of the environment that affords action to appropriately equipped organisms" (p. 118). Affordance theory suggests that a social computing application functions as an "artifact," or more broadly, as an "environment," that offers those features that enable and "afford particular sorts of actions" (p. 185). Affordance is fostered by a social computing application's use of place, a social environment, in contrast to space, its locational characteristics. Thus, affordance theory allows us to understand how the environment and the way it is used play a role in researchers' decisions to use non-library discovery services to obtain OA documents.

According to Dourish (2001), the concept of place leads to several substantial sociological consequences. The first consequence is highlighted by the difference between the two terms *place* and *space*. A place directs our attention away from the environment as simply a structure and towards the environment as a social sphere. Hence, the structure of the surroundings disappear into the background as the space becomes more social. Often then, a "'place' reflects the emergence of practice" (p. 90), and by this Dourish means that a place becomes customized and shaped by its use as we may re-arrange the chairs in a room according to how we use the room. Last, a place may mean different things to any particular community of practice, and so one particular setting may have multiple meanings for any of the communities that use it, and this is dependent upon how they use it.

These insights about social computing provide the necessary framework for understanding how it may shed new insights on scholarly communication. In particular, a social computing application's structure and functionality may afford the tools necessary to create a space where users of that locale converge through a common practice. When these events overlap at a place where the practice concerns scholarly and scientific bibliographic references, the social worlds of authors, scholars, scientists, as well as readers become more accessible to researchers interested in the sociological aspects of scholarly communication as well as the quantitative techniques used to measure it.

2.4.4 Collecting Bibliographic References: Social Computing and Bibliometrics

White and McCain (1989) write that "bibliometrics is grounded in the patterned behavior of human beings---the authors, editors, and indexers on the production side of the world of learned publications. Specifically, it is grounded in the linguistic choices by which they associate indicators of content" (p. 123). They mark a distinction between authors, editors, and indexers on this production side from those on the consumption side, that is the "readers or users." For this consumption side, they divide information science into two categories---information retrieval and information needs and uses.

The online availability of bibliographic records along with the growth in interactive digital libraries, which users help build by providing content, and therefore are also producers in some sense, has resulted in a new blend of these facets of information science. This is where the production and consumption of bibliographic records merges with the authors, editors, and indexers on the publication side and with the readers and users on the consumption side. That is, the readers or users of published scholarly and scientific literature now also produce "the linguistic choices by which they associate indicators of content" with articles and other writings, which are the "true unit of analysis in many bibliometric studies" (White & McCain, 1989, p. 124).

Readers and users contribute to the production side in two significant ways: by selecting, saving, and building second-tier databases of bibliographic records and by tagging them with keywords. The outcome of this activity is the creation of systems, such as CiteULike or Mendeley, that highlight different aspects of information retrieval and information needs and uses as identified by White and McCain (1989). These databases are different from other databases that are traditionally used in bibliometric studies like the ISI indexes, Scopus, and lately Google Scholar. Rather than attempts at storing, organizing, or simply linking to the entirety of scholarly and scientific publications, or some authenticated set of it, these databases (or indexes) are the result of user and/or reader production and therefore consumption-side aggregated value. It is this phenomenon of *readers as indexers* and of what it may reveal about the social world of scholarly communication that is the indirect fuel for this study and of the bibliographic references produced by them that is its object.

It is important to note that users collecting, storing, sharing, and tagging bibliographic references in such web-based social computing applications are not instances of citing behavior. Citing is a "norm" which acknowledges "the work of those who have gone before" (Budd, 1992, p. 348) and citations may be seen, metaphorically, as "signposts" (Smith, 1981, p. 85). In contrast, there is no such permanence involved in adding bibliographic references to online personal, yet public, digital libraries, which may later be deleted. While these bibliographic references do act as a sort of acknowledgment, they do not necessarily act as a sort of acknowledgment in the sense that a citation does, given that they are not situated within published discourse, specifically grounded in argument, or directly serve to promote scientific or scholarly progress based on traditional form of inquiry.

In the sense that these types of social computing applications may act as a type of acknowledgment, they may do so because it is assumed that users of these sites are selecting and adding bibliographic references to documents they might deem to have some value or utility. The assumption implies that users of these sites would not add bibliographic references to publications they deem to have little value or utility given the effort required, albeit minimal but not necessarily trivial, to add references to a personal library. These references may serve as reminders of what to read, to remember, or to use in some other way at a later date. In this respect, the phenomenon that involves both social computing and the collection of bibliographic references might be more akin to collection development in librarianship. However, this analogy would require additional exploration and examination for it to be valid given the serious logistic, political, ethical, and social complexities involved in public and academic library collection management.

If collecting, storing, sharing, and tagging bibliographic references is related to citing behavior in any way, then one purpose of such research might be to explore how this is so. What is most likely occurring though and what is available for exploration is a tier of information use situated between collecting and citing. That is, these applications might provide insight into some intermediate stage between collecting works and later citing them. Therefore, while such research may be able to provide a metric of scholarly communication as well as a tool to help identify emerging trends, it may also provide insights into various kinds of human information behavior. It could very well be that collecting bibliographic references do function within a discourse and, as such, may function as concept symbols (Small, 1978).

Specifically, in comparing citing behavior with collecting behavior, given that both entail bibliographic references to published scholarly or scientific documents, an examination of the assumptions implicit in citation analysis that Smith (1981, pp. 87-89) outlines function as a good starting point for inquiry within the framework of a collection analysis of bibliographic references. That is, the assumptions implicit within citation analysis might be compared to possible assumptions in bibliographic reference collecting and tagging since collecting and citing are conjoined by the same variable, the bibliographic reference.

Therefore, by adapting and modifying Smith's (1981) list of assumptions we should wonder whether 1) collecting a bibliographic reference to a document implies use, or potential use, of that document by the person collecting it; whether 2) collecting a bibliographic reference to a document reflects the merit of that document; and whether 3) users are collecting bibliographic references to the best possible works. With regards to Smith's third point about assumptions in citation analysis, she writes that a number of other factors influence citing behavior and these may include access to the document and awareness of the document. If access to a document is a factor in whether that document gets cited, then an examination of what bibliographic reference types are collected, for instance, in regards to publication models (i.e., open access and traditional) and mode of access (i.e., subscription databases and free search engines), might shed light on this. Furthermore, given that one of the benefits of a social computing application is to share information with others, an examination of the number of times documents are posted by multiple users as well as an examination of how those users became aware of those documents should provide some insight into how information is shared. Such an examination should be quite feasible given that these social computing bibliographic reference applications provide several functionalities to help others become aware of what has been posted.

The fourth assumption in citation analysis Smith outlines regards how documents are related through bibliographic coupling and co-citation. Research into folksonomies and what tags, hashtags, keywords, or terms users apply to bibliographic references should provide insight into how documents relate to each other especially when compared to documents that are actually bibliographically coupled and co-cited in the traditional sense.

The fifth assumption considers whether all citations are equal and the importance of knowing how much weight to apply to a citation. Smith discusses two types of refinements used to judge the weight of a citation: mechanical and intellectual. Again, the tags, hashtags, keywords, or terms used to describe a document by a user might be considered an intellectual refinement if we assume or can verify that the tags used provide some kind of "content analysis" (p. 90). Since research into folksonomic classification, tagging (Kipp, 2011), and hashtagging (Moulaison & Burns, 2012) is active, an analysis of this assumption should be worthwhile.

Such questions may be explored by examining the nature of the bibliographic references users collect, by attempting to determine the various reasons why they are being collected, and by discovering what new assumptions underlie collecting bibliographic references in comparison to citing. Although not directly answered in this study, the following questions about these assumptions are game:

1. Do bibliographic reference managing web-based applications mainly provide a convenient web accessible, locally independent storage for personal collections for users?
2. Do users find value in them because the sites help them discover publications they may not have normally found?
3. Is there some kind of marketing effect in the sharing process, so that when users think about what they add to their collection they think about broadcasting some work; that is, do they consciously think about the possible effect posting a bibliographic reference might have on others?
4. Are users concerned about the quality of their collection and how it reflects upon them as scholars and scientists?
5. Do users post references to articles and other sources that are deemed to be of high quality, perhaps based on journal or author ranking or citation count, or do they post what they simply think is important and useful to their research or education?

White and McCain (1989) write that "bibliometrics can deal only with explicit data" (p. 164). While this remains true, the data provided by bibliographic reference management social computing applications, about what is collected and possibly read by scholars, makes explicit what was previously unavailable in quantitative aggregate. Essentially, the bibliographic references and papers scholars collect may provide new insights in how traditional bibliometric data is used after it has been extracted from subscription databases and the newer, non-traditional, more complicated sources traced and predicted by Cronin (2001).

Alternative explicit data is studied and explored by others. Most importantly, these studies demonstrate an interest in capturing in quantitative aggregate sources of influence, of collaboration, and of recognition not easily identified from a list of references alone. For example, within the journal article, Cronin, Shaw, and La Barre (2003) and Cronin and Franks (2006) pursue an analysis of an article's paratext, its bylines, and its acknowledgments. In webometrics, where the unit of analysis is a web page and the variable under consideration is a hyper link (Thelwall & Harries, 2004; Björneborn & Ingwersen, 2004), interest in supplementing our knowledge of scholarly communication is leveraged by the various technologies implemented socially by scholars and scientists. Vaughan and Shaw (2003) discuss the characteristic differences between web sitations [*sic*] and bibliographic citations and explore these differences in an empirical exploration of influence. Priem (2013) advocates for changes in scholarly communication based on expanded data sources that indicate influence.

Budd (1992) highlights one of the central motivations behind early bibliometrics when he refers to Narin and Moll's 1977 *ARIST* chapter on the subject, in which they say, "many of the early bibliometric papers resulted from an *innate curiosity* [emphasis added] about the functioning of the scientific enterprise" (Narin & Moll, 1977, p. 36, as cited in Budd, 1992, p. 346). The curiosity explored here is how social computing may provide further insights in the area of scholarly communication and the scientific enterprise. In the spirit of this curiosity, this study examines a potentially different sign of influence---the references one collects using a social computing bibliographic reference management application. This study assumes that what academics, scholars, and scientists collect may be as revealing as what they cite, either in overlapping or unique ways.

2.5 Conclusion

While many perceive the purpose of academic libraries to be about collecting, organizing, and providing access to the scholarly record, not all within the profession or the research community agree on the specifics. However, even if collecting, organizing, and providing access to information is the primary purpose of the academic library, the academic library is no longer the sole or primary actor with this function. New sources to discover scholarly information and new publishing models make the academic library a competitor for users and a competitor with users of its services and sources. Valuating the academic library must take these other services and sources into consideration.

This study merges several theories and methodologies in order to answer its research questions. Collecting bibliographic references using bibliographic reference management services such as CiteULike allows us to work with new data types. Although these data exist in the familiar form of a bibliographic reference, they represent an entirely different activity. Rather than being instances of citing, they are instances of collecting, and studying them is possible because of advances in social computing. Using decision and game theory, we can infer from this activity the strategic impact these collecting actions have on academic libraries while still holding some of the assumptions of citation analysis true.

3 PROCEDURES

3.1 Introduction

This study proposes examining the properties of bibliographic references scholars and researchers collect and using a freely accessible bibliographic database to examine additional statistics about these references.

In the first section of this chapter I describe the sources of data, in this case, CiteULike and Google Scholar, which are used for bibliometric and regression analyses. In the next section, I describe the logistic regression method, which is used to determine what predictor variables predict access to full text documents outside of a library's proxy. In the third section, I describe the Bayesian probability method, which will be used with the findings of the Ithaka S+R study (Housewright & Schonfeld, 2010) to determine a hypothetical probability that a library's discovery services and collections were used given an option to use an alternate discovery service and an alternate collection. In the fourth section, I describe the data collection process, and I follow this with a description of the variables used from the CiteULike and Google Scholar data. I finish this chapter outlining the plan of analysis.

3.2 Data Sources

The bibliometric and regression analyses are conducted on data collected from CiteULike and Google Scholar. CiteULike provides the bibliographic references and Google Scholar provides bibliometric and publishing data including citation counts and source of item if the item is freely accessible via Google Scholar without the need of a proxy. Here I briefly describe CiteULike and Google Scholar.

3.2.1 CiteULike

CiteULike has been an object of study and a source of data for studies. It has primarily been used by those interested in folksonomies and tagging (Capocci & Caldarelli, 2008; Kipp, 2011). As of October 2008, less than two years before collecting data for this study, CiteULike.org had “885,310 unique items, annotated by 27,489 users with 174,322 unique tags” (Bogers & van den Bosch, 2008). At least one study used CiteULike, along with two other social bibliographic reference managers, as a source to analyze journal usage (Haustein & Siebenlist, 2011).

CiteULike users may add bibliographic references to their libraries either manually or automatically. In the latter case, adding a bibliographic reference to a personal library is accomplished either via a JavaScript bookmarklet for the browser or through a social bookmarking link on a scholarly document's web page (CiteULike, 2010c). The bookmarklet or bookmarking link will extract bibliographic data from an appropriate web page and import the bibliographic details into its database. Users can assign tags to their references and these will function as a type of "flexible filing system" (Emamy & Cameron, para. 6, 2007). Users may also assign additional metadata, and this includes noting whether the reference refers to the user's own publication (authored), the priority to read the publication, and whether collecting the reference is public or private (default is public) information. Users may also add notes via a simple text editor in the browser and write a review of the publication. Users may view related articles based on the tags that have been assigned by the user adding the bibliographic reference as well as any tag that other users have assigned to the same bibliographic reference. CiteULike will generate a formatted reference in a number of styles including APA, Chicago, IEEE, Harvard, and many others. Finally, users may export their libraries in various formats, either for generating formatted references or for importing into another bibliographic reference manager application.

CiteULike offers a number of social functions. Users may connect with other users and join groups of users who may be interested in similar research or who are working together on a research project. Users can share bibliographic references and write blog entries about those references within the site. Users may also create personal profiles of themselves where they can provide details such as their name, email, location, job title, affiliation, web page, and research fields.

3.2.2 Google Scholar

Google Scholar was introduced in 2004 and has since grown in popularity on several fronts. Research has been conducted on its use and popularity as a search tool among students (Herrera, 2011; Cothran, 2011) and by librarians (Neuhaus, Neuhaus, & Asher, 2008), its ability to index content in institutional repositories (Arlitsch & O'Brien, 2012) or to locate open access content (Norris, Oppenheim, & Rowland, 2008), and its scope (Chen, 2010) and coverage in various subject areas such as history (Kirkwood & Kirkwood, 2011) and engineering (Baldwin, 2009).

Some studies have used Google Scholar as a bibliometric or informetric tool, where the latter methodology refers to a broader notion of bibliometrics and means "the quantitative study of recorded discourse" in any medium (Wolfram, p. 39, 2003). Kousha and Thelwall (2007) compare Google Scholar to the ISI indexes. Noruzi (2005) provides an introduction to Google Scholar's use as a citation analysis tool. Harzing and Wal (2008) describe the use of Google Scholar as a citation analysis tool and offer a free program that uses Google Scholar to compute alternative journal impact scores and other citation measures (see *Publish or Perish* at http://www.harzing.com/pop.htm). Aguillo (2012) conducts a webometric analysis showing that Google Scholar is a problematic source for bibliometrics because its coverage lacks quality control.

Despite Aguillo's (2011) concerns about the quality of sources Google Scholar indexes, Google Scholar is a useful bibliometric tool in this study for two main reasons: 1) because it is used to locate known bibliographic references that have been saved by users in CiteULike; and 2) because Google Scholar functions as the relevant alternative to using the academic library as a research starting point. This study therefore depends on Google Scholar's increased coverage over subscription bibliographic databases such as Scopus and Web of Science since the references that CiteULike users save may themselves be more comprehensive than what the more selective bibliographic databases cover.

Google Scholar offers a number of functions including the ability to locate scholarly works, either through simple or advanced searching, export citations to those works, provide total counts of citations, search within works that cite other works, and link to the full text of works if the full text is available and indexed by Google Scholar. In the latter case, the name of the hostname providing the full text is provided by Google Scholar as a hyperlink to the full text. For example, a full text document with a link to the hostname umsystem.edu would most likely refer to the University of Missouri's institutional repository at mospace.umsystem.edu.

Libraries can use a link resolver to allow Google Scholar to provide access to subscribed content (Google, n.d.). When libraries configure and use this service, Google Scholar seamlessly integrates with the library's collections. This works for the users of a particular library who use Google Scholar within an authenticated Internet Protocol (IP) range, usually that of a university's network. In such cases, it will be necessary for patrons to use Google Scholar on campus or, if off campus, through a virtual private network (VPN) connection.

3.3 Logistic Regression

One of the variables in this study includes whether Google Scholar points to full text copies of the bibliographic references in the CiteULike sample. This variable is a binary or dichotomous data type (Yes / No) and is therefore a candidate as a dependent variable in a logistic regression. Modeling a logisitic regression allows us to test how a set of predictor variables affect or are related to a binary or dichotomous variable (Harrell, 2001). Logistic regression does not assume a normal distribution or linear relationships between the variables (Sin & Kim, 2008). However, a logistic regression requires meeting four assumptions or problems: multicollinearity, independence of errors or cases, linearity of the logit, and no complete separation, which means any one variable should not completely predict any of the other variables (Field, Miles, & Field, 2012); however, separation is generally only a problem when there are multiple categorical or dichotomous variables (Boslaugh, 2012). When the independent (predictor) variables are of the same data type (e.g., ratio), multicollinearity becomes a concern when the predictor variables are highly correlated (Adkins & Bala, 2004; Sin & Kim, 2008). There is no test for independence of errors, which assumes that variables are not related. Testing for the linearity of the logit requires modeling the logistic regression and including an interaction between any continuous predictor variables and the log of itself (Field, Miles, & Field, 2012).

The predictor variables may include both categorical and continuous data (King, 2008), and this study will include the number of authors for each bibliographic reference (author count), the year the bibliographic reference was posted to CiteULike (post year), the publication year of the reference (pub year), and citation counts. The Post Year variable is unique to this study and to a bibliometric analysis and is influenced by the social computing nature of CiteULike. The Publication Year variable can be used to refine the model (Sin, 2011). Based on these variables and the more general theoretical motivations described in this study, the logistic regressions will address whether the variables in the data set predict full text availability in Google Scholar. The regression equation produced by this model should be able to predict which of these variables affect full text availability. The model produces an odds ratio (*OR*) for each of the independent variables in relation to the dichotomous dependent variable. This reflects an overall effect size (Harrell, 2001).

The odds ratio (*OR*) is perhaps the most important statistic, at least for interpretation, resulting from a logistic regression. It is the result of dividing the odds of one group by the odds of a second group and is interpreted by reference to the numerator.

For example, "odds ratios of 2, 0.5, and 1 indicate, respectively, that the odds of the group in the numerator are 100% larger (doubled), 50% smaller (halved), and neither larger nor smaller than the odds of the group in the denominator" (King, p. 366, 2008).

3.4 Bayesian Analysis

The 2009 Ithaka S+R faculty survey (Schonfeld & Housewright, 2010) found that 38% of scientists claim to begin their information seeking with Google, and from that statistic and others like it the authors of that report make the claim that academic libraries are increasingly being disintermediated from the discovery process as a result of this kind of information seeking practice. The problem with that claim is that it does not take into consideration the alternate route. That is, if 38% of scientists use Google as a starting point for their research, then we might say, broadly speaking, that 62% of scientists use the academic library as a research starting point. Although this is a simplification and a broad assumption and the real world choice or sample space is certainly not binary given that it does not take into consideration other discovery mechanisms such as those related to invisible colleges (Price, 1986), the decision between the two represents a near world scenario and contrasting them provides a way to outline the *minima* and *maxima* of the model, or the theoretical upper and lower bounds. Additionally, a set of conditionals is necessary in order to make a claim about the disintermediation of the academic library, and this set of conditionals refers to the success rate of either the academic library or Google Scholar in retrieving a relevant full text document as a result of choosing the academic library or Google Scholar, respectively, as a research starting point. That is, before a valid claim about the disintermediation of the academic library can be made, we must ask the following question: given a research starting point, what is the probability of retrieving a relevant full text document? Thus the meaningful question is, given that 38% of scientists use Google as a research starting point, what percentage of those scientists could hypothetically experience successful retrieval events of relevant documents outside of a university's proxy? Bayes' theorem allows us to invert this question in order to answer the following: what is the probability that a scientist used an academic library (or Google Scholar) as a starting point given having retrieved a full text document. If we can answer that question with some credibility, then we address the claim about the disintermediation of the academic library.

According to Phillips (1973), Bayes' theorem is useful when revising prior opinion or belief in light of new evidence. While it is the basis for more complex Bayesian statistical analysis, Bayesian probabilities can still incorporate frequentist or objective calculations into revised or "judgmental probabilities" (Raiffa, p. 124, 1968). The theorem follows from the three laws of probability, such that if all three laws are accepted, Bayes' theorem must be accepted. These three laws are:

> *First law* Probabilities cannot be less than zero nor greater than one, and the probability of the sure event is 1. Put mathematically,
>
> $0 \leq p(E) \leq 1$ and $p(\text{sure event}) = 1$ (Phillips, 1973, p. 31)
>
> *Second law* The probability of either of two mutually exclusive events occurring is equal to the sum of their individual probabilities. In mathematical notation,

p(E1 or E2) = p(E1) + p(E2) (Phillips, 1973, p. 32)

Third law The probability of both E and F occurring is equal to the probability of E times the probability of F given E. In mathematical notation,

p(E and F) = p(E) x p(F|E) (Phillips, 1973, p. 40)

All three probability laws are standard, but the third law, taking into consideration the joint probability of E and F, directly leads to Bayes' theorem. As Phillips (1973) states: "Bayes' theorem can be obtained by applying the third law to each of the joint probabilities, so that the probability of each joint event is given by the product of an unconditional and a conditional probability" (p. 58). In our case, we have four joint probabilities to consider:

1. The probability of using an academic library and retrieving a relevant full text document;
2. the probability of using an academic library and not retrieving a relevant full text document;
3. the probability of using Google Scholar and retrieving a relevant full text document; and,
4. the probability of using Google Scholar and not retrieving a relevant full text document.

More generally, Bayes' theorem (Phillips, 1973) can be expressed as follows:

$$p(H_1|D)=\frac{p(H_1)\times p(D|H_1)}{p(H_2)\times p(D|H_2)+p(H_1)\times p(D|H_1)} \tag{1}$$

Where:

1. $p(H_1)$ and $p(H_2)$ equal our priors;
2. $p(D \mid H_1)$ and $p(D \mid H_2)$ equal our prior probabilities; and,
3. $p(H_1 \mid D)$ equals our desired posterior probability.

More commonly, we say that given a piece of data D, what is the probability a specific event occurred H_1, or $p(H_1 \mid D)$, where this can mean what is the probability of having used an academic library as a research starting point given having retrieved a relevant full text document. Bayes' theorem does not allow us to compute this without taking into consideration the total data and all the available decisions or events. Such that, we have to know the joint probability of having retrieved a relevant full text document D outside of a university's proxy given having used Google Scholar $p(H_2)$ and the joint probability of having retrieved a relevant full text document D having used an academic library $p(H_1)$. It is not enough to know how successful the academic library is in aiding a searcher in retrieving a relevant full text document without taking into consideration how successful Google Scholar is also, given that these are the two broad options available to researchers, as the Ithaka report claims.

The joint probability we wish to know will be more easily understood with a decision tree. Lindley (1971) describes a decision tree as a "method of analysis" (p. 141) where the simple events are laid out from left to right to illustrate the complex or joint events under consideration, or the product of the unconditional event by the conditional event. As a result, computing the posterior probability is simply a matter of summing what may be considered the "expected payoffs" (Chacko, p. 125, 1991) or the conditional probabilities under examination. This follows from the second probability law outlined above.

3.5 Data Collection

After receiving approval on May 18, 2010 from CiteULike for access to their data, their entire data set was downloaded on May 19, 2010 in two separate files. These files contained identification numbers for each of the references in the CiteULike library and amounted to identification numbers for 2,419,452 unique bibliographic references (CiteULike, 2010a)

These identification numbers were first sorted and deduplicated. In order to acquire a substantial, systematic random sample (Vaughan, 2001; Vaughan & Shaw, 2008), the count of the unique bibliographic references was divided by 1,000. This resulted in the number 2,419. A random number was generated (4,438), and starting at this number, which indicated the 4,438th bibliographic reference in the data, every 2,419th identification number was harvested. This resulted in a sample size of 999 bibliographic references.

Each identification number in the sample was manually used to retrieve the bibliographic reference from the CiteULike web site in its BibTeX format (a format for processing bibliographic references). Four of the 999 references in the sample were missing from CiteULike due to a server error during retrieval. I assume this might be the result of deletion by the user who added the bibliographic reference, but the cause could be a result of some other action or event.

It should also be noted that since CiteULike is a social computing application, users can add academic profiles. I examined 20 CiteULike users with a profile to confirm that users of CiteULike are researchers or scholars of some sort (CiteULike, 2010b). All profiles indicated that CiteULike users were researchers. The sample largely included self-identified graduate/doctoral students, post docs, faculty, and what appear to be corporate or other organizational researchers. They work in a variety of nations including the United States, Germany, Italy, France, the United Kingdom, Denmark, Brazil, and Mexico.

BibTeX is a machine and human readable format for bibliographic references and is commonly used in the science and mathematical disciplines along with LaTeX, a document formatting programming language. Depending on the bibliographic reference, the BibTeX format may contain most of the relevant information including: Document type, Abstract, Address, Author, Publication date, DOI, ISSN, ISBN, Journal, Keywords/tags, Volume, Issue, Pages, Posted date, Title, and URL. I used this information to find the sources in Google Scholar, but for data analysis, I only include the Author count, Publication date, Posted date, and Document Type. The full document type list includes: Book, Book chapter/section, Booklet, Collection (part), Conference inproceedings (part), Conference proceedings (whole), Electronic citation, Journal article, Manual (technical documentation), Miscellaneous, Technical report, Thesis (Master's), Thesis (PhD), and Unpublished work.

Using Google Scholar, I collected three years of bibliometric and publishing data.

The data was collected on July 14, 2010, on July 17-19, 2011, and on July 14-16, 2012. Google Scholar was used to collect data on the following variables: Found (yes/no), Citation Count, Full Text Access (yes/no), and Full Text Source. Notes were kept for any items that seemed inconsistent. For example, some of the bibliographic references referred to simple web pages and there were some instances when Google Scholar found a citation one year but not the next. Also, I conducted the search outside of the university's proxy or network. This insured that full text sources, outside the subscription pay wall, are truly full text sources. However, not all links were tested and it is possible that some of these links were broken. This is a limitation of the study.

3.6 Description of Variables

The data sources are CiteULike and Google Scholar. CiteULike provides the initial data set of bibliographic references. The variables from CiteULike include:

1. Document Type: Includes the type of document found in the sample of bibliographic references. This includes the common formats: journal articles, proceeding articles, and books.
2. Posted Year: The year the bibliographic reference was posted to CiteULike by a CiteULike user.
3. Published Year: The year the bibliographic reference indicates the source was published.

I use Google Scholar to examine the bibliographic references. The variables from Google Scholar include:

1. Citation Count: The number of citations Google Scholar shows for each

bibliographic reference.

2. Found: This variable indicates whether Google Scholar was able to find the bibliographic reference and return a link or a citation to it. The result is either true or false.
3. Full Text Access: Whether Google Scholar was able to find a full text copy of the source. We use the term full text and not open access because we do not make any assumptions about the licensing status of the document. The result is either true or false.
4. Full Text Source: If a full text document was found for the bibliographic reference, this indicates the source providing the full text. Such sources may include institutional repositories, open access journals and databases, academic portfolio web sites, pre-print archives, or others.

3.7 Plan of Analysis

The analysis begins with a description of the overall sample. The majority of the sample of bibliographic references point to the journal article document type, and for the sake of measurement consistency, I perform most of the analysis on this document type. This analysis includes how many of the bibliographic references were found by Google Scholar. I then proceed to show how many journal articles Google Scholar provides full text access to. Since I collected data on the sources providing full text access and in order to show the state of decentralization of the storage of scholarly information, I show the most popular individual sources providing such access. These sources are then classified by type, such as university, governmental, and publisher, and I provide a breakdown of the type of sources providing full text access via Google Scholar. This follows with a standard bibliometric analysis, which includes showing the publication date distribution of the articles, the posting date distribution, the citation counts, and for comparison, the citation counts of those articles that Google Scholar is able to provide a link to a full text source against those it is not able to provide such a link. This comparison illustrates the influence of what is collected, such that if full text is more influential (i.e., has higher average citation counts) than non full text, then we should consider this a factor influencing whether a researcher collects an item (its relevance). This influence is then tested with the logistic regression in order to model what is predicting full text availability. Finally, I end the analysis using Bayes' theorem to assess the hypothetical probability that the academic library or Google Scholar was used as a research starting point.

The bibliographic references collected from CiteULike were saved in a spreadsheet file. Data collected from Google Scholar was added to this file under additional columns. The data was then cleaned and exported to a comma separated value (CSV) file and imported into RStudio (http://www.rstudio.com/), an integrated development environment (IDE) for the R programming language (R Core Team, 2012). The R programming language was used for the analysis along with several packages that extend its functionality. These packages include *ggplot2* (Wickham, 2009), *reshape2* (Wickham, 2007), and *lubridate* (Grolemund & Wickham, 2011). All software used is free and open source software.

4 RESULTS

4.1 Introduction

The purpose of this analysis is to determine whether the academic library is being disintermediated by researchers' information discovery processes and the decentralization of scholarly content, and consequently, risks marginalization. It is certainly true that scholarly information seekers have a vast amount of tools available to them to conduct their queries and a vast amount of options available to them from where they can retrieve their documents. Since not all these services or collections are provided by the academic library, as was the case for much of the 19th and 20th centuries, the data analyzed in this chapter should shed light on the impact these services and sources have on the current state of affairs.

4.2 Bibliometric Analysis

4.2.1 CiteULike's Coverage

CiteULike users appear to collect a great variety of document types including journal articles, books, proceeding articles, and so forth. However, some document types are more abundantly collected than others. As seen in Table 1, a majority of the sample is to the article document type (69.45%), and this is followed by references to books (8.94%) and proceeding articles (8.94%). Since the article document type dominates the sample, and because issues with open access largely concern journals (although not necessarily), much of the analysis will focus on the references to articles.

Table 1

CiteULike Sample Composition

Document Type	***Count***	***Percentage***
Article	691	69.45%
Book	89	8.94%
In Proceedings	89	8.94%
Misc	39	3.92%
Electronic	18	1.81%
Proceedings	17	1.71%
In Collection	15	1.51%
Tech Report	15	1.51%
PhD Thesis	9	0.90%
In Book	6	0.60%
Unpublished	3	0.30%
Master's Thesis	2	0.20%
Booklet	1	0.10%
Manual	1	0.10%
Total	995	99.99%

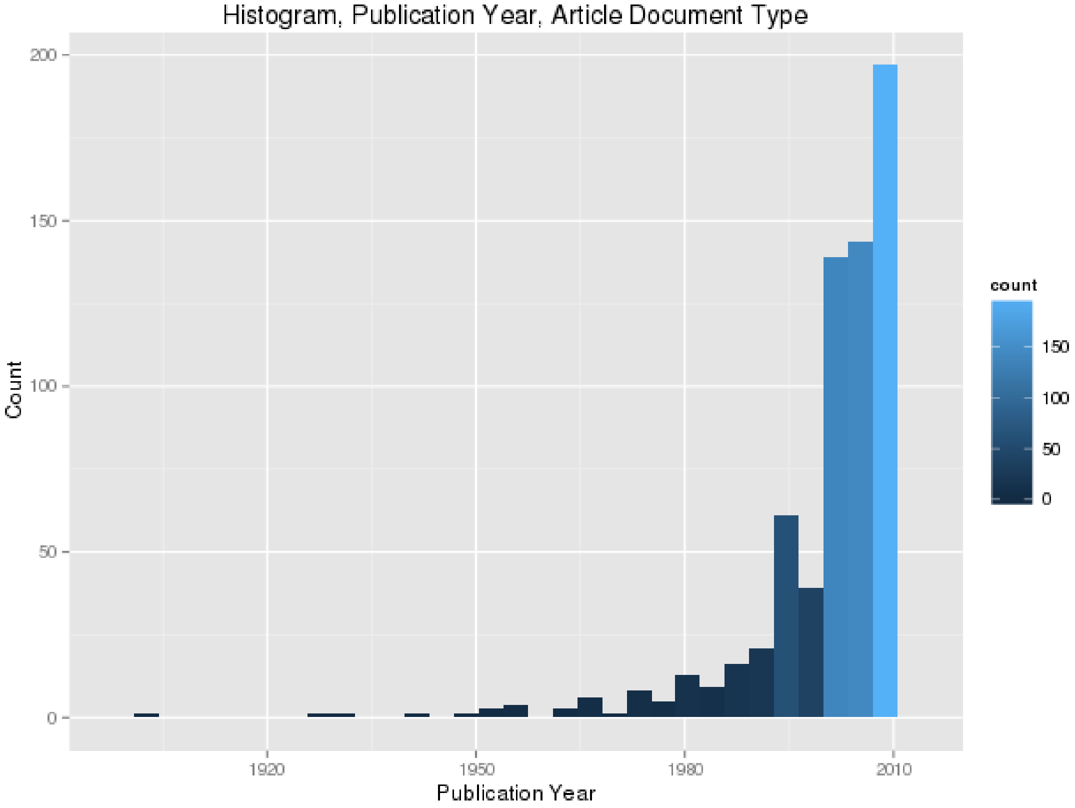
CiteULike users collect articles that have been published over the last hundred years, but as represented by the median publication dates in the sample of articles in Table 2, most of what is collected has been published in the latter part of the 20th century and the first part of the 21st century. The post date of the articles illustrates the usage of CiteULike, indicating the middle value between the founding of CiteULike in 2005 and when the source data was collected in 2010.

Table 2

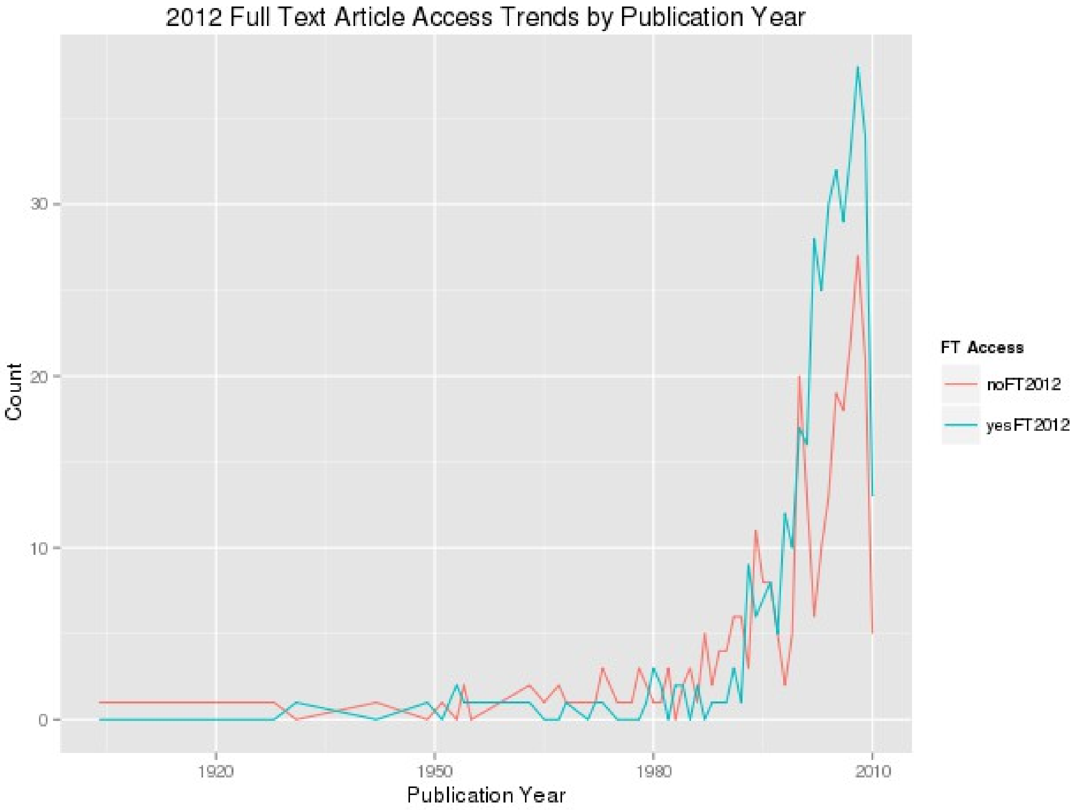
Publication and Posting Years for Articles

Date Variables	***n***	***Mdn***	***Min***	***Max***	***NAs***
Pub Year	674	2004	1904	2010	17
Post Year	691	2008	2005	2010	0

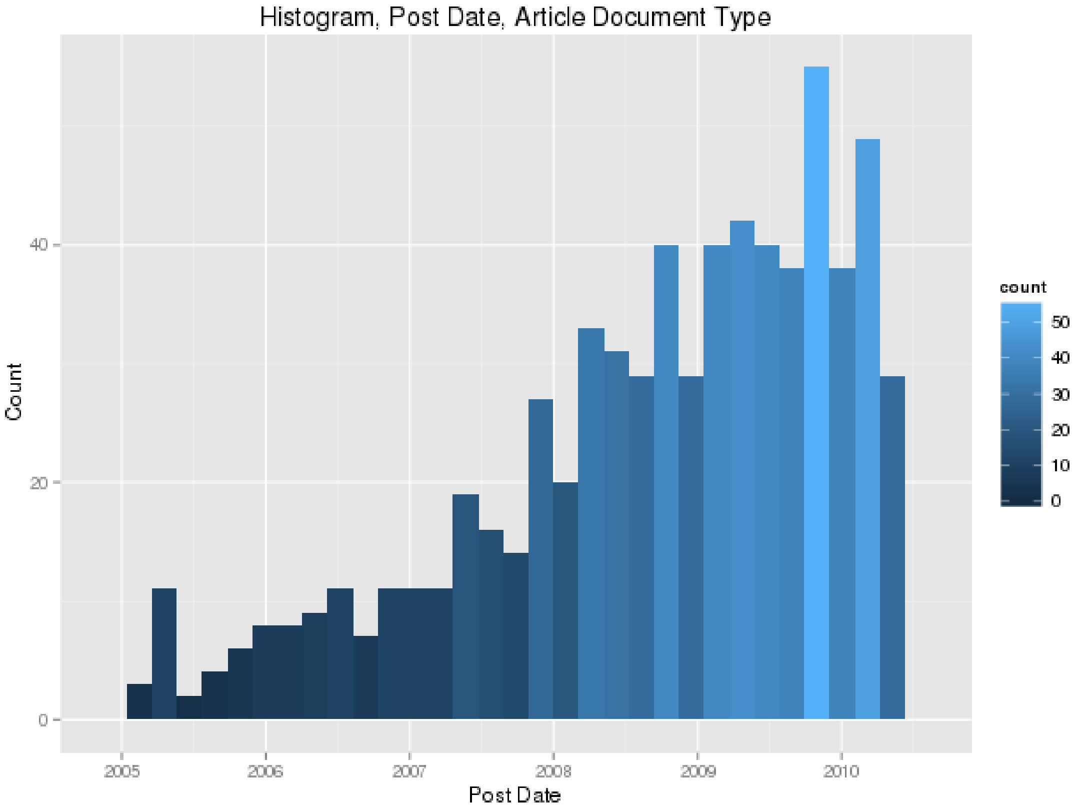
Figures 1 and 2 show the distribution of articles by *publication year* and the collecting trend over this time frame. While the sample shows some interest in articles published as early as the beginning of the 20th century, most articles that have been collected have been published since the year 2000 (Figure 1). The trend lines in Figure 2 highlight the tendency to collect articles that were available via Google Scholar in 2012, especially those articles that were published more recently. Likewise, Figures 3 and 4 show the distribution of articles by *posting year* and the collecting trend for this time frame. Figure 3 highlights the growing use of CiteULike, as of mid-2010 when the sample was taken, and Figure 4 shows that CiteULike users always tended to collect articles that would potentially become freely or openly accessible.



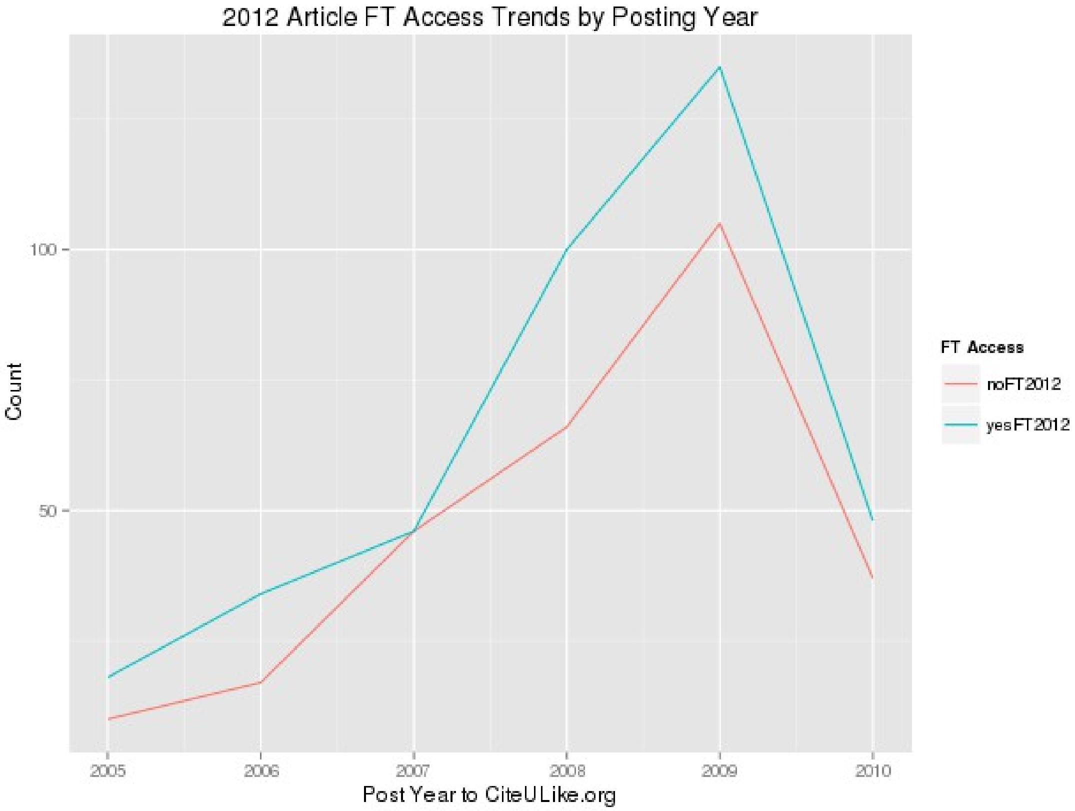
1. Figure: Histogram of publication years of articles collected by CiteULike users



2. Figure: Trend lines showing that CiteULike users collect articles that tend to be free or open access more than not via Google Scholar.



3. Figure: Histogram of posting years of articles collected by CiteULike users.

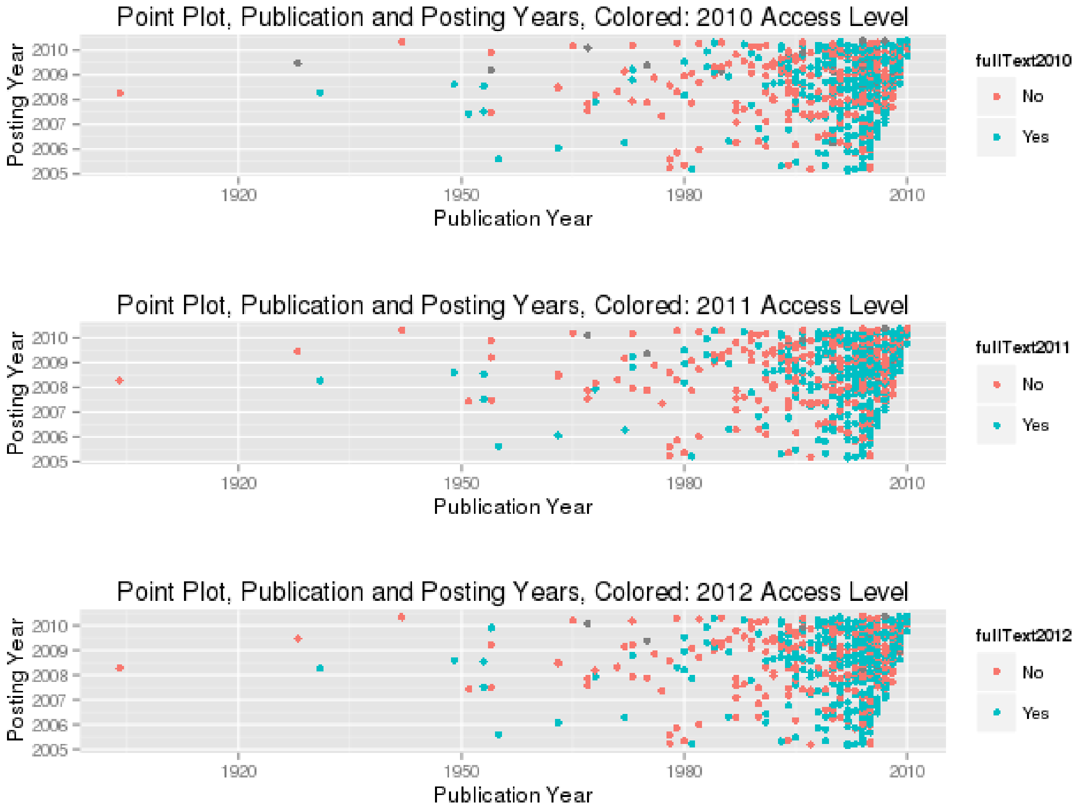


4. Figure: Trend lines showing that CiteULike users collect articles that tend to be free or open access more than not via Google Scholar.

4.2.2 Google Scholar's Coverage

Since the validity of Google Scholar as a bibliographic database is pertinent to this study, it is important to know how well Google Scholar can locate known items. In 2010, Google Scholar was able to locate 648 out of the 691 references to journal articles. The reasons all 691 of the references were not located vary, but the list of reasons are that some references point to items besides journal articles, such as to news articles, even though users or the CiteULike system have classified them as article types. In other cases, the bibliographic references are in a foreign languages (Chinese, etc.), and as such I cannot confirm their search hits in Google Scholar. In other cases, the bibliographic references are incomplete and cannot be verified. Despite this, in 2010, I was able to locate 648 of the 691 bibliographic references in Google Scholar, and this increased to 663 in 2011 and saw a very slight drop to 662 in 2012.

Figure 5 shows the three years of distributions of article types by publication year and posting year, with colored dots representing full text access (blue) or no full text access (red). As is evident from the plots, full text access dominates non full text access via Google Scholar.



5. Figure: Publication year by posting year collecting trends. Points represent full text access and indicate that CiteULike users tend to collect items that are or become free or open access.

4.2.3 Full Text Access

Table 3 shows the full text breakdown by year. Controlling for the relative yearly increases in the bibliographic references that were discoverable through Google Scholar, the increase in full text access from 2010 (345 / 648) to 2012 (381 / 662) is 8.10%. A one-sample proportions test with continuity correction was completed for each of the yearly data where the null hypothesis is that the probability of the proportion equals 0.5 (c.f., Newcombe, 1998). Neither of the 2010 full text access variables (No and Yes) were significantly different from 0.5 at the 95% confidence level, and this is illustrated by the overlap between the upper bound confidence interval for the 2010 *No* variable and the lower bound confidence interval for the 2010 *Yes* variable. However, as the annual availability of access to journal articles increases, the alternate hypothesis that the probability of the proportion does not equal 0.5 becomes more acceptable. By the year 2012, when 381 out of 662 articles, or over 57%, were found by Google Scholar, the spread between the upper bound confidence interval of the 2012 *No* variable (46.32%) and the lower bound confidence interval for the 2012 *Yes* variable (53.68%) increases substantially from that in 2010, where there was a statistical overlap resulting in a non-significantly different measure (i.e., upper *CI* for 2010 No Full Text is 50.69% crosses the lower *CI* for 2010 Yes Full Text which is 49.31%). Essentially, holding a sample of bibliographic references to articles constant, the probability that we will be able to retrieve a full text copy from Google Scholar, without the benefit of a university's proxy, substantially increases by 2012 for the sample.

Table 3

Article Count with Google Scholar Full Text Access, 2010 – 2012

	Full Text	***Count***	***Estimate***	χ^2	***df***	***p***	***Lower CI***	***Upper CI***
2010 (*n* = 648)	No	303	46.76%	2.5941	1	0.1073	42.87%	50.69%
	Yes	345	53.24%	2.5941	1	0.1073	49.31%	57.13%
2011 (*n* = 663)	No	299	45.10%	6.178	1	0.0129	41.28%	48.98%
	Yes	364	54.90%	6.178	1	0.0129	51.02%	58.72%
2012 (*n* = 662)	No	281	42.45%	14.8051	1	0.0001	38.66%	46.32%

	Full Text	***Count***	***Estimate***	χ^2	***df***	***p***	***Lower CI***	***Upper CI***
	Yes	381	57.55%	14.8051	1	0.0001	53.68%	61.34%

4.2.4 Full Text Sources

The number of full text articles that are freely available through Google Scholar appears to be a function of the number of unique sources providing full text access. In 2010, 176 unique sources provided full text access to 345 articles via Google Scholar. In 2011, the number of unique sources increased to 190 and these sources provided access to 364 of the articles in the sample. In 2012, 229 unique sources provided access to 381 articles. Overall, this represents a 29.94% increase in the number of unique sources providing full text access, from 2010 to 2012, and a 8.10% increase in the full text articles that are available, after controlling for differences for each year's total sample. Dividing these numbers by the three year time period implies that for every 9.98% point increase in the number of unique sources, there is a 2.70% point increase in the number of full text articles that are available. Thus, as scholarly sources of information become more decentralized and grow in number, the probability that full text material (e.g., open access articles) is accessible outside of a university's proxy increases.

Table 4 lists the most frequent unique sources providing full text access to more than three articles in the total sample of articles. For example, in 2010, Google Scholar linked to CiteSeerX to provide the majority of full text access to articles, but by year 2012, Google Scholar linked to CiteSeerX to provide full text access to just five articles in the top list of full text source providers. The remaining unique sources hold fairly steady across the time period. Lastly, four of the top ten sources reference full text articles under the Green OA publishing model while six link to full text articles under the Gold OA model.

Table 4

Full Text Article Sources with More than Three Instances: 2010 – 2012

Full Text Source	***2010***	***2011***	***2012***	***OA Type***
CiteSeerX	40	38	5	Green
NIH	35	42	40	Gold
arXiv	27	28	26	Green
Oxford Journals	12	13	12	Gold
PNAS	11	11	11	Gold
BioMed Central	7	10	11	Gold
PLoS	5	4	5	Gold
Harvard University	5	5	5	Green
Rockefeller University	4	--	4	Green
American Meteorological Society	--	--	4	Gold

All sources providing full text access to the articles in the sample were classified by type according to their hostname name, each of which was visited.[1] Appendices A, B, and C provide a comprehensive list of the counts, hostnames, and classifications I used for these tables. Classifying these sources involved subjective decision-making. For example, I classified NIH.gov as a *Government* source but I classified France's

[1] The hostname was visited but not the full text document. That is, when Google Scholar provides a hyperlink to a full text document, the name of the hyperlink is simply the hostname plus the top-level domain. For example, umsystem.edu has the hostname *umsystem* and the top-level domain name *.com*. Since I only visited the hostname and not the full text document, some inferences about the classification I used had to be made. For example, that umsystem.edu refers to an institutional repository, such as the one at mospace.umsystem.edu and not a research group's web site or some departmental web site located at another subdomain of umsystem.edu.

multidisciplinary open archive *HAL* (http://hal.archives-ouvertes.fr/) as a *National* source. If the source was affiliated with a university, unless it was a personal academician's web site, and thus classified as *Personal*, I classified the source as a *University* source. Sources in the *University* class include institutional repositories, some subject repositories that are operated by universities or university libraries (e.g., arXiv.org), and departmental or research group sites. For-profit and non-profit journal publishers were classified as *Publisher*. If the source was affiliated with an academic or professional association, such as the American Psychological Association, I classified it as a *Publisher*. One web site with a social justice mission (fahamu.org) provided full text access to one document for the 2012 year. In order to maintain consistency, all sources for all three years of data were classified at the same time, in mid-January 2013.

The classification suggests that universities, and primarily institutional and subject repositories, remain important points of access for full text documentation. Table 5 provides a breakdown of the unique sources providing full text access. Most significantly, universities account for 56.82% of the unique sources providing full text access to articles in 2010. In year 2012, for these same articles, this increases by over six percentage points to 63.32%. All percentages have been rounded to the second decimal spot.

Table 5

Full Text Sources by Type, Count and percentage of unique source types providing full text access

Type	***2010***	***2011***	***2012***
Activism	--	--	1 (0.44%)
Business	7 (3.98%)	8 (4.21%)	10 (4.37%)
Government	4 (2.27%)	4 (2.11%)	4 (1.75%)
National	3 (1.70%)	3 (1.58%)	5 (2.18)
Organization	1 (0.57%)	--	1 (0.44%)
Personal	5 (2.84%)	7 (3.68%)	9 (3.93%)
Publisher	40 (22.73%)	40 (21.05%)	46 (20.09%)
University	100 (56.82%)	117 (61.58)	145 (63.32%)
Unknown	16 (9.09%)	11 (5.79%)	8 (3.49%)
Sum	176 (100%)	190 (100%)	229 (100%)

Table 6 provides a breakdown of the number of documents each unique source is providing access to. Although it could be true that a small number of unique source types provide access to a majority of the documents, it does not hold true here. For example, although government agencies only account for a small percentage of the unique source types providing full text access, it would be possible that this source type provides a large percentage of the documents. However, the data suggest varied relationships for many of the cases. For example, Tables 5 and 6 show that in 2010, four unique government sources provided full text access to 39 articles and 100 university provide access to 183 articles in 2010.

Table 6

Full Text Source by Type, Count and percentage of number of articles each unique type of source is providing access

Type	***2010***	***2011***	***2012***
Activism	--	--	1 (0.26%)
Business	7 (2.03%)	8 (2.20%)	11 (2.88%)
Government	39 (11.30%)	46 (12.64%)	46 (12.04%)
National	5 (1.45%)	5 (1.37%)	6 (1.57%)
Organization	1 (0.29%)	--	1 (0.26%)
Personal	5 (1.45%)	7 (1.92%)	9 (2.36%)
Publisher	88 (25.51%)	87 (23.90%)	100 (26.18%)
University	183 (53.04%)	200 (54.95%)	199 (52.09%)
Unknown	17 (4.93%)	11 (3.02%)	9 (2.36%)
Sum	345 (100%)	364 (100%)	382 (100%)

4.2.5 Citation Analysis

Citations, as counted by Google Scholar, show fairly substantial increases over the three year time period. Here the median and not the mean is a much stronger description of the middle location of the distribution. This is consistent with previous studies of citation distributions (e.g., Vaughan & Shaw, 2008; Vieira & Gomes, 2010). Table 7 shows that the median citation count for the sample of articles was 23 in 2010, which increased by five median points for 2011 and by an additional nine median points by 2012. The high *Max* value for the citations illustrates the non-normal distribution of citation counts.

Table 7

Article Citation Counts: Years 2010--2012

Year	***n***	***Mdn***	***Min***	***Max***	***NAs***
2010	648	23	0	6,156	43
2011	663	28	0	7,062	28
2012	663	37	0	8,374	28

The distribution of citation counts for the three years follows a fairly traditional citation distribution where a greater percentage of articles receive a relatively fewer number of citations and a smaller percentage of articles receive a relatively greater number of citations. However, the distribution does not follow a consistent 80/20 split, such as is often hypothesized by Bradford's Law (Brookes, 1969). Still, the citation distributions do follow a common inverse relationship (Wolfram, 2003), even if the proportions are different.

Table 8 is organized by quartile division of cumulative percentage of articles and shows their distributions for the years 2010, 2011, and 2012. A reading of the table indicates, for example, that for 2010, 162 articles (25.00%) have a citation count less than or equal to four, or 0.02% of all citations. Holding the cumulative percentage of articles constant, for 2011 we see that 167 articles (25.19%) have a citation count less than or equal to seven, or 0.04% of all citations. Furthermore, for each of the three yearly measures, around 75% of the articles receive between four and five percent of the citations such that a little less than 25% of the articles receive approximately 95% of the citations. This suggests that most of the references to articles that CiteULike users collect may be considered low to moderately influential, with respect to citation counts.

Table 8

Distribution of Articles and Citation Counts, ordered by quartile division of the cumulative percentage of articles (3rd Column)

	Cumulative Sum of Articles	***Cumulative Percentage of Articles***	***Citation Count***	***Cumulative Percentage of Citations***
2010	162	25.00%	4	0.02%
	327	50.46%	23	0.47%
	490	75.62%	72	4.27%
	648	100.00%	6,156	100.00%
2011	167	25.19%	7	0.04%
	334	50.38%	28	0.58%
	499	75.26%	83	4.70%
	663	100.00%	7,062	100.00%
2012	166	25.04%	11	0.07%
	333	50.23%	37	0.79%
	500	75.41%	102	5.04%
	663	100.00%	8,374	100.00%

Table 9 illustrates the previous point and shows the distribution when ordered by the quartile divisions of cumulative percentage of citation counts. Here we read the table such that, for example, 25.11% of all citations have a citation count less than or equal to 285 and belong to 597 or 92.13% of the articles. This time holding the cumulative percentage of citations constant, in 2011 25.44% of all citations have a citation count less than or equal to 348 and belong to 612 or 92.31% of the articles.

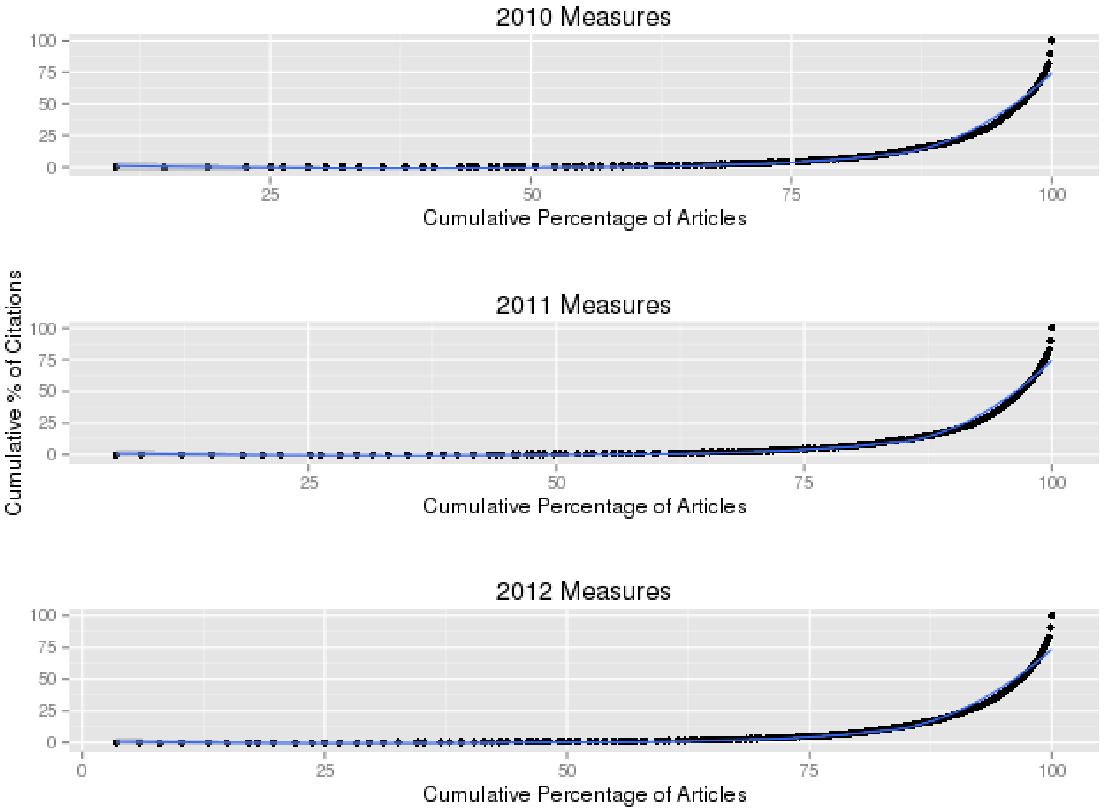
In Table 8, for the year 2010, around 75.62% of the articles in the sample account for only 4.27% of the citation counts, and in Table 9 we see that 76.38% of the citation counts account for 99.38% of the articles and have a citation count of 1,591 or less. In essence, the majority of articles that CiteULike users collect have very few citations in proportion to the highly cited articles that CiteULike users collect. If citations are a measure of influence, then it seems that a majority of what CiteULike users collect are articles that have low impact.

Table 9

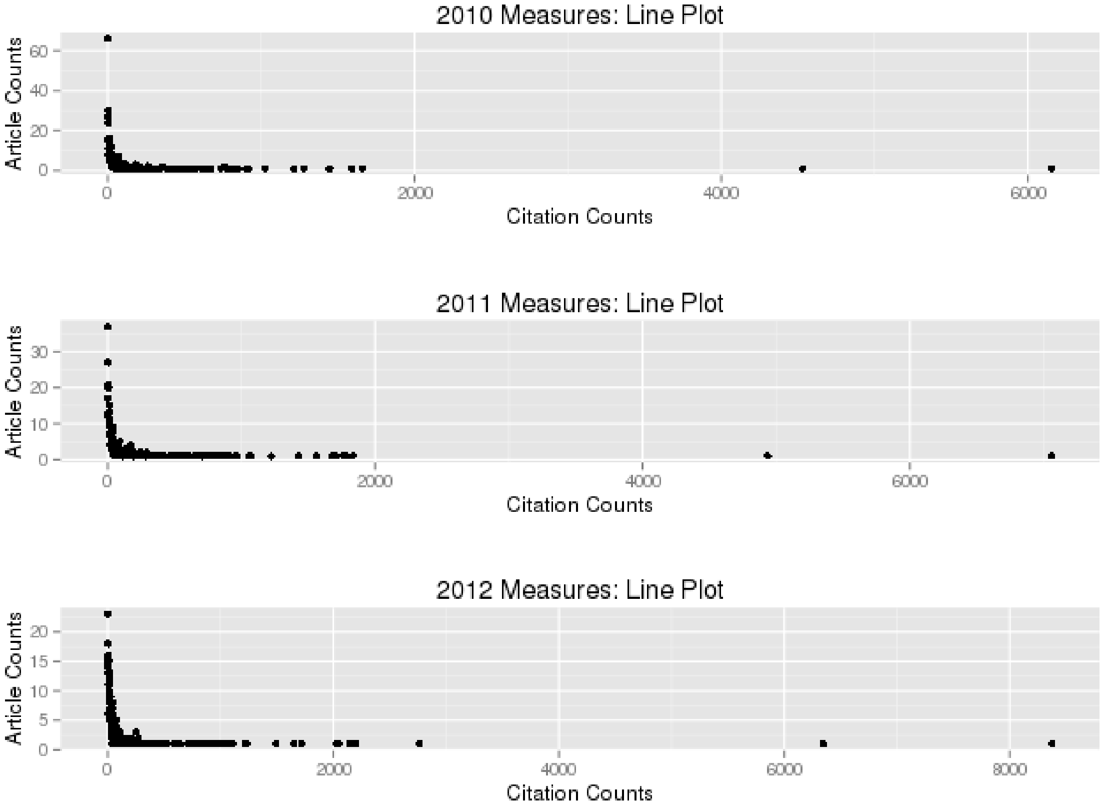
Distribution of Articles and Citation Counts, ordered by quartile division of the cumulative percentage of citations (3rd Column)

	Cite Count	***Cumulative Percentage of Citations***	***Cumulative Sum of Articles***	***Cumulative Percentage of Articles***
2010	285	25.11%	597	92.13%
	736	50.24%	628	96.91%
	1,591	76.38%	644	99.38%
	6,156	100.00%	648	100.00%
2011	348	25.44%	612	92.31%
	838	51.08%	643	96.98%
	1,702	75.22%	658	99.25%
	7,062	100.00%	663	100.00%
2012	372	25.05%	605	91.25%
	937	50.85%	642	96.83%
	2,145	75.51%	658	99.25%
	8,374	100.00%	663	100.00%

Figures 6 and 7 graph these distributions. Figure 6 shows that for all three years, 75% of the articles account for less than 5% of the citations. Figure 7 shows that as citation counts increase, the number of articles with high citation counts is fewer. Again, the vast majority of articles have very few citation counts.



6. Figure: For all three years, a majority of the articles account for less than 5% of the citations, despite the rapid growth in citation counts for each of the years.



7. Figure: For all three years, most articles have low citation counts.

There is a citation difference between articles that are available full text via Google Scholar and articles that are not available because they may be behind a pay wall. Although we know from Table 2 that there is no statistically significant difference between full text availability of article counts for the 2010 measures, as shown in Table 10, there is a substantial difference between median citation counts for the 2010 full text availability of articles. Specifically, articles that were referenced in the CiteULike sample and that were not full text available via Google Scholar in the year 2010 had a median citation count of 12 compared to a median citation count of 32 for articles that were available.

Table 10

Article Citation Counts by Full Text Access: Years 2010--2012

Year	***Full Text***	***n***	***Mdn***	***Min***	***Max***
2010	No	303	12	0	1,662
	Yes	345	32	0	6,156
2011	No	299	15	0	1,833
	Yes	364	37	0	7,062
2012	No	281	20	0	2,048
	Yes	381	49	0	8,374

This 20 point spread increases to a 22 point spread between the median citation counts of non-full text and full text available articles for the year 2011 and a 29 point spread for the year 2012. The spread from year 2010 to 2011 represents a 10 percentage change, and from year 2011 to 2012 the spread represents a 31.82 percentage change. While the spread appears to be increasing, the median count of the non-full text available articles appears to be increasing at a faster rate of change than the median count of the full text available articles. For non-full text available articles, the median citation count for 2011 is 15 and for 2010 it is 12. This represents a 25 percentage increase. For 2012, the median citation count of non-full text available articles is 20, and this represents a 33.33 percentage change from the year 2011. For full text available articles, the median citation count for 2010 is 32 and the median citation count for 2011 is 37. While the spread is greater, in absolute value, than non-full text available articles, this represents only a 15.63 percentage change, which is a much slower percentage change than the 25 percentage change for non-full text article median count. For full text available articles, the median citation count in 2012 was 49. This represents a 32.43% rate of change from the year 2011, which is slightly slower than the comparable yearly difference for non-full text available articles. Additional years will have to be collected before suggesting any certainty that the rates of change are stabilizing or converging. It does, however, suggest agreement with other open access citation studies that argue that open access articles have an advantage in the initial lead, but part of this advantage may be attributed to articles coming out from behind pay walls, if for example they were NIH funded and were made open access at a later date (National Institutes of Health, n.d.).

4.3 Logistic Regression

Although it appears that what is influencing full text availability via Google Scholar outside of a university's proxy is both the number and type of sources providing full text availability, the citation difference between full text and non-full text documents and the dispersion and growth over the three years suggests there might be a positive relationship between higher citation counts and full text availability. In order to test whether citation counts predict full text availability, plus other variables that might be a factor, I use logistic regression to model these influences.

The logistic regression models the influence of several predictor variables on a dichotomous dependent variable. The predictor variables include Author Count, Publication Year, Post to CiteULike year, and Citation count. The dependent variable is the binary outcome of the full text value: Yes equals full text available and No equals full text not available. Although three years of citation data were collected, because high citation counts may indicate the following year's full text availability, I only model two logistic regressions in order to determine whether the citation counts for 2010 journal articles predict their full text availability in 2011 and whether the citation counts for 2011 journal articles predict full text availability in 2012.

Tables 11 and 12 show the summary statistics for both logistic regressions. All assumptions have been met[2] and both models show that they are better than chance at predicting the outcome (Field, Miles, & Field, 2012). The latter was assessed by computing the statistical significance of the chi square distribution of the null and residual deviances of the models. Both have a *p*-value of less than 0.

Table 11 shows the predictor variables on the 2011 full text availability variable. The odds ratios for the Author Count, the Publication Year, and the 2010 Citation count are statistically significant. The Post year is not statistically significant. This means that it has no statistical influence on the availability of access to full text articles via Google Scholar. The logistic regression model is:

$$\begin{aligned} Logit(p) = 120.2 + 0.0887(Authors) + 0.0425(Pub\,Year) \\ -0.1023(Post\,Year) + 0.0015(Citations\,2010) \end{aligned} \tag{2}$$

[2] The author count variable for the 2012 regression (Table 12) violates the linearity of the logit at the $p =$ 0.10. We include it in the model since the threshold is a 95% confidence level. In any case, removing it has no real influence on the model.

Table 11

Logistic Regression on Full Text Dichotomous Variable: 2011 Article Full Text Access, with Exponentiated Coefficients and Confidence Intervals

Variable	***B***	***SE***	***Wald t***	***Prob***	***Lower (95%)***	***OR***	***Upper (95%)***
Authors	0.0887	0.0318	2.795	0.0052**	1.0301	1.0928	1.1665
Pub Year	0.0425	0.0101	4.201	0.0000***	1.0238	1.0434	1.0653
Post Year	-0.1023	0.0646	-1.582	0.1136	0.7946	0.9028	1.0241
Citations2010	0.0015	0.0005	2.984	0.0028**	1.0006	1.0015	1.0025

Significance Codes: 0 `***`   0.001 `**`

Table 12 shows the predictor variables on the 2012 full text availability variable. This time the odds ratios for Author Count and Post Year are not statistically significant but the Publication Year and the Citation counts for 2011 are statistically significant.

Table 12

Logistic Regression on Full Text Dichotomous Variable: 2012 Article Full Text Access, with Exponentiated Coefficients and Confidence Intervals

Variable	***B***	***SE***	***Wald t***	***Prob***	***Lower (95%)***	***OR***	***Upper (95%)***
Authors	0.0100	0.0198	0.503	0.6148	0.9761	1.0100	1.0583
Pub Year	0.0473	0.0098	4.828	0.0000***	1.0294	1.0484	1.0697
Post Year	-0.0911	0.0644	-1.415	0.1571	0.8038	0.9129	1.0350
Citations2011	0.0016	0.0005	3.345	0.0008***	1.0007	1.0016	1.0025

Significance Codes: 0 `***`   0.001 `**`

The logistic regression model for Table 12 is:

$$\begin{gathered} Logit(p)=88.4637+0.0100(Authors)+0.0473(Pub\,Year) \\ -0.0911(Post\,Year)+0.0016(Citations\,2011) \end{gathered} \tag{3}$$

To determine the influence of the statistically significant variables, the odds ratio can be used to calculate the difference between variables at different points (Boslaugh, 2012). For example, the *OR* for 2011 full text author count is 1.0928, which suggests that the more authors an article has, the more likely the article will be available full text. To compute the predicted change between an author count of one and an author count of five, I take the difference of five and one and place it in the exponent for the odds ratio of the author count: $1.0928^4 = 1.4261$. Thus, the predicted change in the odds of an article with an author count of five compared to an author count of one is 1.4261. Although citation counts have a much greater range than author counts, the influence will be controlled by the relatively neutral odds ratio for the 2010 citations counts. Consider the predicted change for an article with a citation count of 101 compared to an article with a citation count of one: $1.0015^{100} = 1.1617$.

The predicted probability can be found by converting the logits of the two models (Boslaugh, 2012). The predicted probability is computed by the following equation:

$$Predicted\ probability = \frac{e^{(logistic\ regression equation)}}{\left(1 + e^{(logistic regression\ equation)}\right)} \qquad (4)$$

Table 13 summarizes the predicted probabilities (see Appendix D for the computations). In essence, when all variables are held constant at the first quartile mark, the 2011 model suggests there is 49.59% probability that the article will be available full text through Google Scholar outside of a university's proxy. This increases by nearly five percentage points for the 2012 model. When the values are held constant at the third quartile mark, the predicted probability increases substantially. In the 2011 model, there is a 60.82%

probability that an article will be available full text and 63.34% probability it will be available full text in 2012. Although not all the odds ratios are statistically significant for each model, and caution is advised before accepting them wholesale, the models do suggest that as each variable increases in count, the probability that an article will become available full text increases over time.

Table 13

Summary of Predicted Probabilities of Full Text Access for 2011 and 2012 Logistic Regression Models

Range	***2011 Model***	***2012 Model***
First Quartile	49.59%	54.06%
Median	56.27%	59.84%
Third Quartile	60.82%	63.34%

4.4 Bayesian Hypothetical

The data collected from CiteULike and Google Scholar tell us the probability of retrieving a full text document based on those documents that have been identified as relevant to the users that have collected them. In other words, if we assume that the bibliographic references collected by CiteULike users represent documents that are relevant to them, and since we can determine how many of those documents can be retrieved from Google Scholar, then we can infer the success rate of Google Scholar. That is, we can infer the probability of retrieving a full text article given having used, hypothetically, Google Scholar as a research starting point. For scientists, as of 2010 and according to Ithaka S+R (Schonfeld & Housewright, 2010), this is 38%. Given this, what follows is a hypothetical exploration. While it is not a statistical analysis, the exploration provides a heuristic or an intuitive way of thinking about the impact of alternate discovery services and decentralized and openly accessible scholarly content on academic libraries. That this is the current state of affairs represents a degree of uncertainty about how and where scholars search for and retrieve scholarly content.

Bayesian probability allows for the ability to make an educated guess about a set of conditionals given the two broad options for the research starting points. The process is outlined by Phillips (1973) and proceeds first by selecting two hypotheses:

H_1: *Use academic library as research starting point.*

H_2: *Use Google Scholar as research starting point.*

And adding notation for marking the outcome of either:

D_1: *The data marking the retrieval of a full text document.*

D_2: *The data marking the non-retrieval of a full text document.*

The next step is to identify the prior probabilities. As reported, the 2009 Ithaka S+R faculty survey (Schonfeld & Housewright, 2010) provide the statistic for the use of Google as a research starting point (38%) among scientists, and from that statistic, I broadly infer the complement, the probability that the academic library was used as a research starting point. Thus, the probability of the priors, or the probability of the first hypothesis and the probability of the second is: $p(H_1) = 0.62$ and $p(H_2) = 0.38$.

Using Ithaka's statistic and the statistics identified in this study concerning the success rate of Google Scholar, I determine the probability of retrieving a full text document D_1 given having hypothetically used Google Scholar $p(H_2)$ as a research starting point. As shown in Table 2, Google Scholar located 57.55% of the sampled articles for 2012. For simplicity's sake, I round this up to 58%. Applying the third law of probability, I multiply 38% by 58% to get the conditional probability $p(D_1 \mid H_2)$ of having collected a full text document given having used Google Scholar as a research starting point. Thus the $p(D_1 \mid H_2) = 0.2204$ or 22%. Likewise, I calculate the failure or non-retrieval D_2 given having used Google Scholar as a research starting point. Thus, the $p(D_2 \mid H_2)$ is 0.38 times 0.42 or 0.1596 or 16%.

Although I know of no contemporary studies that show how many relevant, known documents an academic library can retrieve, I give the academic library the benefit of the doubt and attribute to it a score of 97%. That is, I suppose that an academic library can acquire 97% of the articles in the CiteULike sample and can do so either through its collection on hand, from its collection in storage, from its subscribed content, or through inter-library loan. Earlier in the study, I stressed the point that if scientists use Google Scholar as a research starting point 38% of the time, then the complement must be broadly true and they use the academic library 62% of the time. While this is a simplification of the sample space and does not consider other potential research starting points, such as practices resulting the existence of invisible colleges and the use of social media, it emphasizes the reality that using the academic library as a research starting point has a maximal upper bound. Thus, the probability of having used the academic library as a research starting point *p*(H_1) is 0.62 and the probability of retrieving a full text copy of one of the articles *p*(D_1) is 0.62 times 0.97. This means that the *p*(D_1 | H_1) is 0.6014 or 60%. Likewise, I calculate the failure or non-retrieval D_2 given having used the academic library as a research starting point. Thus, the *p*(D_2 | H_1) is 0.62 times 0.03 or 0.0186 or 2% after rounding.

In sum, the above calculations provide the total set of prior probabilities that are needed to compute the desired posterior probabilities, where the posterior probabilities are the probability that a CiteULike user used the academic library if she collected a full text document for her article bibliographic reference and the probability that a CiteULike user used Google Scholar if she collected a full text document for her article bibliographic reference. In notation with descriptions, my posterior probabilities are:

> *p(H_1 | D_1) : The probability that a CiteULike user used the academic library as a research starting point if she collected a full text document for her bibliographic reference.*

> *p(H_2 | D_1) : The probability that a CiteULike user used Google Scholar as a research starting point if she collected a full text document for her bibliographic reference.*

To compute these posterior probabilities, I complete Bayes' theorem with the details above. Thus, where Bayes' theorem is:

$$p(H_1|D)=\frac{p(H_1)\times p(D|H_1)}{p(H_2)\times p(D|H_2)+p(H_1)\times p(D|H_1)} \tag{5}$$

Then:

$$p(H_1|D_1)=\frac{(0.62\times 0.6014)}{((0.38\times 0.2204)+(0.62\times 0.6014))}=0.8165=82\% \tag{6}$$

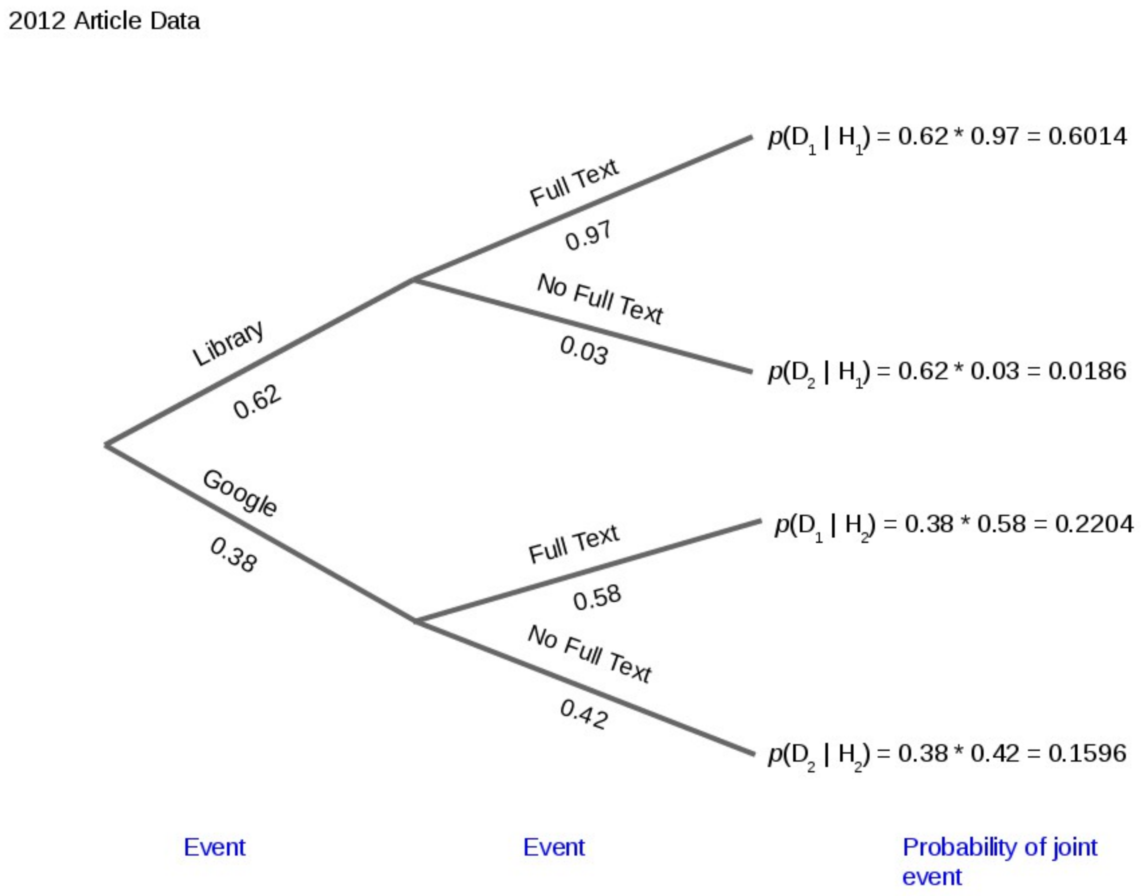
And:

$$p(H_2|D_1)=\frac{(0.38\times 0.2204)}{((0.38\times 0.2204)+(0.62\times 0.6014))}=0.1834=18\% \tag{7}$$

Consequently, I revise the prior information from the Ithaka study (Schonfeld & Housewright, 2010) with the new information from this study to make the following two claims, after rounding:

1. There is an 82% maximal probability that a CiteULike user used the academic library as a research starting point if she collected a full text document for her article bibliographic reference.
2. There is an 18% minimal probability that a CiteULike user used Google Scholar as a research starting point if she collected a full text document for her article bibliographic reference.

Figure 8 shows the above in the form of decision three. Computing the joint probabilities is simply a matter of working left to right for each fork. The result is the four prior probabilities, which are used to find the posterior probabilities.


2012 Article Data
Library
0.62
Google
0.38
Full Text
0.97
No Full Text
0.03
Full Text
0.58
No Full Text
0.42
$p(D_1 \mid H_1) = 0.62 * 0.97 = 0.6014$
$p(D_2 \mid H_1) = 0.62 * 0.03 = 0.0186$
$p(D_1 \mid H_2) = 0.38 * 0.58 = 0.2204$
$p(D_2 \mid H_2) = 0.38 * 0.42 = 0.1596$
Event
Event
Probability of joint event

8. Figure: Decision tree outlining the conditional probabilities that build into Bayes' theorem.

4.5 Conclusion

This chapter began with a bibliometric analysis of a systematic random sample of data collected from CiteULike and augmented by data collected from Google Scholar. I began with an overview of the entire sample and then proceeded to focus on the article document type. This was done for the sake of measurement consistency and also because the article document type is the most popular document type in the sample and open access issues largely pertain to journal articles. I then showed that Google Scholar was able to provide full text access to a majority of the articles in the sample. While the proportion was not significantly different in the year 2010, it was by the year 2011 and more so by the year 2012 because of the increasing number of articles collected in the 2010 sample that became full text available. Although the sources providing full text access via Google Scholar are varied, when classified by type I show that the dominant source providing full text access to journal articles is the university, which should be largely composed of two sources: institutional and subject repositories.

The bibliometric analysis of the article type, by publication date, by post date, and by citation count show that the articles exhibit fairly typical characteristics with those in other bibliometric and citation counts. This weakly suggests that CiteULike users are not very different from scientists in general, an important consideration in inferring the composition of the CiteULike population. A surprising finding was that those articles with full text availability via Google Scholar showed a rather substantial citation advantage compared to those articles that were not full text accessible via Google Scholar. This supported the notion that citations might be a factor of full text availability.

In order to determine what factors influence full text availability, I conducted two logistic regressions using a selection of predictor variables that might show full text availability. The two logistic regression models provided overall fits and the predicted probabilities derived from the models suggest an interesting influence on full text availability; however, statistically significant variables shifted between the two years.

Although this warrants additional modeling, the results suggest that the main influence lies outside the variables tested.

Lastly, I used Bayes' theorem to build a hypothetical probability profile that would infer the likelihood of the academic library's use. This profile drew upon a statistic found in the Ithaka S+R 2009 faculty survey report (Schonfeld & Housewright, 2010) that showed that 38% of scientists report the use of Google as a research starting point. Adding that number with the data from this study, I drew two inferences about the use of both Google Scholar and the academic library given the possibility of having retrieved a relevant full text document to an article bibliographic reference in the sample. These inferences are:

1. There is an 82% maximal probability that a CiteULike user used the academic library as a research starting point if she collected a full text document for her article bibliographic reference.
2. There is an 18% minimal probability that a CiteULike user used Google Scholar as a research starting point if she collected a full text document for her article bibliographic reference.

If we suppose that a CiteULike user is like any scientist (i.e., from comparable populations), then these claims may generalize to the broader scientific community, although further testing is needed before too many generalizations can be drawn.

Based on the analysis, this study suggests that what predicts full text availability is simply the number of sources providing full text access to articles. As these numbers increase, so does the number of accessible full text articles. Based on the classification of sources providing full text access to articles, in 2012 we know that universities (e.g., institutional or subject repositories, largely) provided 52.09% of the documents in the article sample (see Table 5). When this takes into consideration the Bayesian hypothetical assessment, not only is there an 82% maximal probability that a CiteULike user used the academic library as a research starting point if she collected a full text document for her article bibliographic reference, but over half of the articles she might have retrieved if she used Google Scholar as a research starting point come from a university source. This result has strategic implications for academic libraries, which will be discussed in the following chapter.

5 DISCUSSION AND CONCLUSION

5.1 Introduction

Early in this study, I cited literature showing that researchers increasingly use alternate discovery services as research starting points slightly less or as often as the services provided by academic librarians. Furthermore, since open access content is retrievable by these search engines or other alternate discovery services and since the amount of open access content is growing, then it is likely that many researchers can fulfill much of their informational needs by retrieving open access content with these tools. Similar reasoning has led to the claim that academic libraries will become marginalized by these information seeking practices.

This study applied decision theory and bounded rationality to frame this claim. I showed how it is rational to begin with an alternate discovery service such as Google Scholar when it is possible to retrieve relevant scholarly documentation. I used three years of bibliometric data based on a systematic random sample of bibliographic references collected by users on a social bookmarking web site to measure how many of the bibliographic references are found by Google Scholar and refer to freely available scholarly articles outside of a university's proxy. One key finding was that in 2012, nearly 58% of the bibliographic references to journal articles were freely available from 229 unique sources but that academic libraries provide over half of this content, possibly either through subject or institutional repositories. I also showed that the number of academic libraries providing access to these journal articles have also increased over the three year time period. Given the success of these tools and the growing amount of material available as OA, researchers act rationally no matter which of the two broad choices they make to begin their research starting point and when researchers can still access relevant material.

The dominance of the university in providing full text access to material when researchers use Google Scholar as a research starting point is the piece of evidence that has the strongest impact on the strategic future of the academic library. Collectively, it implies that academic librarians' use of institutional repositories to provide open access content appears to be serving them well (Burns, Lana, & Budd, 2013). The larger implication, though, comes from generalizing the strategic response that institutional repositories specifically serve. This is, access to collections should not be dependent on the popular information seeking practices of any specific population. Rather, they should be inherently flexible and be able to meet, without much or any intervention, whatever information seeking practices are in use.

5.2 Discussion about Strategy

The two main research questions in this study inquire into the claim that academic libraries are being disintermediated or becoming marginalized by the availability of alternative discovery services and by the increased decentralization of scholarly information. While the specific claim made by the Ithaka S+R report is one of the most recent of these claims, the claim itself is not new even though the present state of affairs perhaps gives it renewed import.

The claim itself is based on the idea that one of the academic library's core functions is to collect scholarly information. The implicit argument is that if academic libraries have competitors in the collection “business,” and if the use of their collections is being challenged by these competitors, then academic libraries risk marginalization. Accepting this definition of academic libraries and this argument as it stands, this study shows that even though the storage of scholarly information has become decentralized, we can infer that academic library collections continue to be used to access scholarly information, despite the research starting point. We can therefore reject the premise that others have made about the marginalization of academic libraries.

It may make rational sense for a scientist or any researcher to use a non-library electronic discovery service such as Google Scholar. If it takes less effort to use such a service, and if that service does its job well, then such activity apparently satisfices and is therefore rational under bounds. That rationality must be emphasized in any strategic interaction between librarians and their users or potential users. Still, librarians seem to be responding appropriately by providing open access content, either in the form of subject or institutional repositories, that can be retrieved through alternative services. While using a third party discovery service to retrieve open access or freely accessible content is a relevant alternative to the library's services, i.e. those that it pays for, librarians continue to insert their activities by providing content through open access archiving. The relevant alternative, that is, using Google Scholar or the like, thus appears quite challenging, but librarians seem to be, in aggregate, responding in a competitive fashion.

Librarians have at least three types of competitors. The first type includes those who provide alternate collections, the second type includes those who provide the discovery tools to search for and retrieve those collections, and the third type includes the information seekers. A simple heuristic supports these claims but can also be used to compose strategic plans. This heuristic can be framed as: given the actions taken by agent A, what is the strategic response that maximizes the outcome and equilibrates the game and where the domain of A may include the three types of competitors listed above. If the actions and the agents are relevant to the mission and purpose of the responder, then the heuristic applies. When this heuristic is not used, either by those who make claims for or against academic libraries' role in the scholarly communication system, problems arise. All too often, these claims are based on the idea that new technologies, new players, and new practices will simply by their existence threaten the use of the library.

These claims are simplistic when they do not take into consideration the relevant alternatives or the conditional likelihoods of choosing these alternatives. In this context, it is not appropriate to value a thing in and of itself. It is only appropriate to value a thing compared to a similar thing and to do so iteratively. Measuring the value of an academic library must then take into consideration measuring the value of comparable entities who provide similar services and tools and whose services and tools are used for similar tasks. For instance, when the Ithaka S+R report showed that 38% of scientists use Google as a research starting point and then make the argument that this, and other similar findings, implies the marginalization of the academic library, they fail to highlight the complement, that possibly up to 62% of scientists use the academic library as a research starting point. More pointedly, they also fail to inquire into the conditionals, that 38% of the scientists who use Google as a research starting point may be drawing over 50% of their scholarly content from collections provided by the academic library. The academic library is not being disintermediated; rather, the system is simply growing more complicated and interconnected.

Academic librarians do have challenges in front of them. If discoverability and access to their collections are dependent on the use of specific applications, then academic librarians cannot succeed in responding strategically to the popular information seeking practices of the day. As stated in the beginning of this study, such a scenario is not fully capable of taking into consideration the decision matrix of the information seeker. Currently, for instance, online public access catalogs (OPAC) contain their bibliographic records in the deep web, where the content is primarily discoverable only by using the OPAC search application. Consequently, there is generally only one main path to identify that item in the collection, and that one main path is dependent on the use of a specific tool. Limiting access in such a way is a poor strategic response to the information practices that are common today among users of any classification. In this way, the library is threatened with a disintermediation of the *use* process, if not also the search process. The problem is well known and it could be that current efforts underway to grow the Digital Public Library of America (DPLA) will resolve this issue by using a platform that allows libraries to coordinate in such a way that access to collections are search tool agnostic (see Peek, 2012 for a description of the DPLA).

Despite that academic librarians are responding competitively to the more varied ways and from the more varied locations that researchers search for and retrieve scholarly documentation, academic librarians may still face a competitive disadvantage if researchers do not see or believe that the materials they collect, read, and use do not come from academic libraries when they do. Researchers' subjective beliefs about the costs and payoffs attached to various search strategies may be skewed towards the branding associated with alternate search and retrieval routes and sources, and this skew may reinforce the invalidity inherent in incomplete information about the role that academic librarians play in the scholarly communication system. In essence, librarians might want to continue to respond to the fact that many researchers hold, and act upon their, subjective beliefs about which search strategies result in the least average rate of probable work (Zipf, 1949), but also recognize that, given what we understand about bounded rationality, these researchers are rational agents.

Furthermore, while the open access movement is important for librarians and their communities, the availability of freely accessible, relevant scholarly material does represent a challenge to what an academic library collection means and how academic librarians define the library and themselves. That is, the open access movement does represent numerous advantages for many scholarly stakeholders, but it also represents an existential shift for academic libraries and for the role and profession of librarianship. It is now impossible for academic librarians to exercise "completeness and control" (Smith, 1990, p. 9) of the scholarly record, and this state of the affairs suggests rather significant implications for the library and the profession.

5.3 Future Research

However the future of collection development and management works in practice, it has always been a false argument that the academic library is defined solely by its collection building. Academic libraries are, in fact, defined by the librarians who work in them. What makes the library more than a warehouse of content is the people and the profession that gives that library purpose and intent. Although Plutchak (2012) argues that the future of libraries is librarians, this has long been the case; it is just time, as Plutchak argues, to recognize that. Indeed, as Lingel (2012) writes, albeit on a different topic, the "... Library reflects the values of its community through its policies, not through its collections" (Policies are politics section, para. 1). Hill (2009) notes, also on a different topic, that "Policies guide the organization and the responsibility to create them confers a great amount of power to the creator" (p. 87). These policies, it is important to observe and within the context of this study, are a reflection of the intent of the librarians who write them.

Since this study was able to identify a list of universities that provide open access archived content, future research will involve extensive qualitative research with the people at these institutions in order to inquire into the strategic nature of their policies and practices. The guiding research question in such studies would be: what inherent strategic qualities exist in the practices and policies of those who make a library more than a warehouse of material?

Additionally, Ellis (1989) and Ellis, Cox, and Hall (1993) propose an information seeking model composed of a variety of features that outline the information seeking practices of social and physical scientists. This model's amended form, designed by Meho and Tibbo (2003), is a relevant extension of this study and will form a research design based on email interviews with CiteULike, or comparable, users. This future project will also incorporate questions designed to measure researcher preferences about research starting points.

Although scholarly social computing applications may help researchers identify topical material, a theoretical understanding of what this means with respect to relevance, as pursued in information science, is not well understood. The important assumptions outlined by Smith (1981) about citation analysis apply. Mizarro (1997; 1998) outlines most of the more perplexing issues with relevance. It will be a matter of time before we know how collecting or even tweeting a bibliographic reference indicates anything about the reference's relevance to another user or how, as an example, it might function as a concept symbol (Small, 1978), if it does at all.

5.4 Conclusion

Bibliometrics and information seeking studies both aim to understand information behavior using two different approaches. The former furthers our understanding about general patterns of behavior while the latter offers methods for gaining deeper understanding of the various personal dimensions of the seeking and gathering processes. Using one to build on the other is a complimentary process as using various demographic studies may be used to further our knowledge of specific groups of people through in-depth qualitative inquiry. Additionally, the availability of personal collections of reading material offers an attractive means for inquiring into both the scholarly communication system and the information seeking and gathering behavior of researchers. However, this study has focused less on overall behavior and concentrated more on the inherent decisions of information seekers and their strategic outcomes. Because of this focus, this study highlights the rationality of these decision makers.

The material used in this study provided a rich source of data for understanding how context influences, constrains, and binds such behavior. This material provided important insights about the decisions users make when searching for and saving scholarly content. Lastly, the study used these methodologies and theories to understand the impact that various alternatives have on academic libraries. This impact on academic libraries has largely been ignored or when it has been addressed, it has been studied by applying incomplete premises that have lead to incomplete conclusions. Future inquiry into the future of academic libraries should always take into consideration the entirety of the system and not focus on the isolated actions of any set of people.

APPENDIX A: 2010 Source Classification

2010 sources providing full text to articles with counts and classification of sources

Count	***Source***	***Type***
1	128.131.166.46	university
1	130.102.44.245	publisher
2	135.196.210.195	publisher
1	148.204.64.201	personal
1	59.to	unknown
3	aacrjournals.org	publisher
1	ahajournals.org	publisher
1	ajcn.org	publisher
2	annals.org	publisher
1	anu.edu.au	university
1	apa.org	publisher
1	archives-ouvertes.fr	national
1	arizona.edu	university
27	arxiv.org	university
1	asb.dk	university
1	asu.edu	university
1	berkeley.edu	university
2	biologists.org	publisher
7	biomedcentral.com	publisher
1	birdflumanual.com	business
1	blit.li	unknown
1	bmj.com	publisher
1	brown.edu	university
1	bu.edu	university
1	caltech.edu	university
1	cam.ac.uk	university
1	cancer.gov	government
1	cc.ia.us	unknown
3	cell.com	publisher
1	cjb.net	unknown
1	confex.com	unknown
1	corgentum.com	business
1	cship.org	unknown
1	cshlp.org	business
1	dbkgroup.org	university
1	dicp.ac.cn	university
1	digitalhumanities.org	organization
1	dtic.mil	government
3	duke.edu	university

Count	***Source***	***Type***
1	dur.ac.uk	university
1	econmsu.org	university
2	ejbjs.org	publisher
1	embl-ebi.ac.uk	unknown
1	e-moka.net	unknown
1	emory.edu	university
1	epfl.ch	university
1	fgv.br	university
1	fh-vorarlberg.ac.at	university
1	francetelecom.fr	unknown
1	gel.org.br	university
3	genetics.org	publisher
1	genomebiology.com	publisher
5	harvard.edu	university
1	heatherlench.com	personal
1	helsinki.fi	unknown
2	hematologylibrary.org	publisher
1	hindawi.com	publisher
1	iadis.net	publisher
1	ias.ac.in	publisher
1	idei.fr	university
1	idep-fr.org	university
1	ijbnpa.org	publisher
1	illinois.edu	university
1	infn.it	university
1	infonortics.eu	unknown
1	innovatieforganiseren.nl	business
1	intmedpress.com	publisher
1	irit.fr	university
1	ismni.org	publisher
1	itcj.edu.mx	university
1	ithaca.edu	university
1	iub.edu	university
3	jbc.org	publisher
1	jhu.edu	university
1	jmu.edu	university
2	jneurosci.org	publisher
1	joplink.net	publisher
1	ktu.edu	university
1	le.ac.uk	university
1	lebedev.ru	university
1	letunic.com	personal
1	lincoln.ac.uk	university

Count	***Source***	***Type***
1	lodz.pl	unknown
1	lth.se	university
1	lyellcollection.org	publisher
1	manchester.ac.uk	university
1	mcponline.org	publisher
2	miami.edu	university
1	mit.edu	university
2	mpg.de	university
1	mvr1.com	personal
1	nanofemtolab.qc.ca	university
1	ncsu.edu	university
1	nd.edu	university
35	nih.gov	government
1	nrc-cnrc.gc.ca	national
1	ntnu.no	university
1	ntu.edu.tw	university
1	nyu.edu	university
1	otago.ac.nz	university
1	ovgu.de	university
2	ox.ac.uk	university
12	oxfordjournals.org	publisher
1	persoenlichkeitspsychologie-potsdam.de	university
2	petra.ac.id	university
1	physiology.org	publisher
1	physoc.org	publisher
1	plantcell.org	publisher
3	plosjournals.org	publisher
11	pnas.org	publisher
1	pnexpert.com	personal
2	princeton.edu	university
40	psu.edu	university
1	psycnet.org	publisher
1	psykiatriskforskning.dk	university
1	qualcomm.net	unknown
1	rbej.com	publisher
1	rei.edu	university
1	rhbnc.ac.uk	university
2	rockefeller.edu	university
2	royalsocietypublishing.org	publisher
1	rsna.org	publisher
4	rupress.org	publisher
1	rutgers.edu	university
1	sagebrush.com	business

Count	***Source***	***Type***
1	sagepub.com	publisher
1	santafe.edu	university
3	scielosp.org	national
1	sdsu.edu	university
1	sfu.ca	university
2	sgmjournals.org	publisher
2	shouxi.net	unknown
1	slidearts.com	business
1	soton.ac.uk	university
1	ssji.net	unknown
2	stanford.edu	university
1	sunysb.edu	university
1	toronto.edu	university
1	ttu.edu	university
1	tue.nl	university
1	uci.edu	university
1	uconn.edu	university
1	ucsd.edu	university
1	ucsf.edu	university
1	uea.ac.uk	university
1	ufrgs.br	university
2	ugr.es	university
1	umd.edu	university
1	umich.edu	university
1	unam.mx	university
1	uni-bonn.de	university
1	unifi.it	university
1	uni.kl.de	university
1	uni-muenchen.de	university
1	unlp.edu.ar	university
1	uoregon.edu	university
1	uq.edu.au	university
1	uran.donestsk.ua	business
1	usask.ca	university
2	usda.gov	government
1	usenix.org	publisher
1	ust.hk	university
1	uta.edu	university
1	u-tokyo.ac.jp	university
3	utoronto.ca	university
1	uv.es	university
1	victoria.ac.nz	university
1	washington.edu	university

Count	***Source***	***Type***
2	wisc.edu	university
2	wustl.edu	university
1	wvu.edu	university
1	wwu.edu	university
1	yale.edu	university
1	ym.edu.tw	university
1	yorku.ca	university
1	zuom.info	unknown

APPENDIX B: 2011 Source Classification

2011 sources providing full text to articles with counts and classification of sources

Count	***Source***	***Type***
1	128.131.166.46	university
1	148.204.64.201	personal
1	203.189.120.190	unknown
1	210.45.114.81	unknown
1	aacrjournals.org	publisher
1	ahajournals.org	publisher
1	ajcn.org	publisher
1	alphamedpress.org	publisher
3	ametsoc.org	publisher
2	anu.edu.au	university
1	apa.org	publisher
1	archives-ouvertes.fr	national
1	arizona.edu	university
28	arxiv.org	university
1	asb.dk	university
1	aspetjournals.org	publisher
1	asu.edu	university
1	aut.ac.ir	university
1	benthamscience.com	publisher
2	biologists.org	publisher
10	biomedcentral.com	publisher
1	birdflumanual.com	business
1	birjournals.org	publisher
1	bris.ac.uk	university
1	brown.edu	university
1	bu.edu	university
1	cam.ac.uk	university
1	cancer.gov	government
1	chronobiology.ch	university
1	cjb.net	unknown
1	corgentum.com	business
1	craigmcclain.com	personal
1	cw.com.tw	publisher
1	cyganiak.de	personal
1	dbkgroup.org	university
1	dicp.ac.cn	university
1	dkmic.de	university
1	dtic.mil	government
2	duke.edu	university

Count	***Source***	***Type***
1	dur.ac.uk	university
1	econmsu.org	university
1	emis.de	publisher
1	e-moka.net	unknown
1	emory.edu	university
1	epfl.ch	university
1	fgv.br	university
1	fhv.at	university
1	francetelecom.fr	unknown
1	free.fr	unknown
1	fsu.edu	university
1	gel.org.br	university
1	genego.com	publisher
3	genetics.org	publisher
1	google.com	unknown
1	griffith.edu.au	university
5	harvard.edu	university
1	hawaii.edu	university
1	heatherlench.com	personal
1	helsinki.fi	unknown
2	hematologylibrary.org	publisher
1	hi.is	university
1	hindawi.com	publisher
1	iadis.net	publisher
1	ias.ac.in	publisher
1	idei.fr	university
1	idep-fr.org	university
1	illinois.edu	university
1	infn.it	university
1	intmedpress.com	publisher
1	irit.fr	university
1	ismni.org	publisher
1	itcj.edu.mx	university
1	ithaca.edu	university
1	iub.edu	university
3	jbc.org	publisher
1	jbjs.org	publisher
1	jee.org	publisher
1	jhu.edu	university
1	jmu.edu	university
3	jneurosci.org	university
1	joplink.net	publisher
1	ktu.edu	university

Count	***Source***	***Type***
1	langers.nl	unknown
1	latech.edu	university
1	le.ac.uk	university
1	letunic.com	personal
1	lincoln.ac.uk	university
1	liu.se	university
1	lth.se	university
1	lyellcollection.org	publisher
1	manchester.ac.uk	university
1	marquette.edu	university
1	mcponline.org	publisher
1	miami.edu	university
1	mit.edu	university
1	mpg.de	university
1	mshri.on.ca	university
1	mvr1.com	personal
1	nanofemtolab.qc.ca	university
1	nber.org	business
1	nb.rs	national
1	ncsu.edu	university
42	nih.gov	government
1	ntnu.no	university
1	ntu.edu.tw	university
2	nyu.edu	university
1	otago.ac.nz	university
1	ovgu.de	university
1	ox.ac.uk	university
13	oxfordjournals.org	publisher
1	persoenlichkeitspsychologie-potsdam.de	university
2	petra.ac.id	university
1	physiology.org	publisher
1	physoc.org	publisher
4	plos.org	publisher
11	pnas.org	publisher
1	pnexpert.com	personal
38	psu.edu	university
1	psykiatriskgrundforskning.dk	university
1	qualcomm.net	unknown
1	rachel.org	business
1	rclis.org	university
1	rei.edu	university
1	rhbnc.ac.uk	university
2	rockefeller.edu	university

Count	***Source***	***Type***
3	royalsocietypublishing.org	publisher
1	rsna.org	publisher
3	rupress.org	publisher
1	rutgers.edu	university
1	sagebrush.com	business
1	sagepub.com	publisher
1	santafe.edu	university
3	scielosp.org	national
1	sdsu.edu	university
2	sgmjournals.org	publisher
1	shouxi.net	unknown
1	slu.se	university
3	stanford.edu	university
1	sunysb.edu	university
1	swarthmore.edu	university
1	syr.edu	university
1	toronto.edu	university
1	tribler.org	business
1	tsukuba.ac.jp	university
1	tuc.gr	university
1	tue.nl	university
1	ucalgary.ca	university
1	uci.edu	university
1	ucla.edu	university
1	uconn.edu	university
1	ucsc.edu	university
1	ucsd.edu	university
1	uea.ac.uk	university
2	ugr.es	university
1	ukpmc.ac.uk	university
1	umassmed.edu	university
1	umb.edu	university
1	umd.edu	university
2	umich.edu	university
1	uned.es	university
1	Uni-dortmund.de	university
1	unifi.it	university
1	uni.kl.de	university
1	uni-muenchen.de	university
1	unl.edu	university
1	uoa.gr	university
2	upenn.edu	university
1	uran.donestsk.ua	business

Count	***Source***	***Type***
1	usc.edu	university
2	usda.gov	government
1	usenix.org	publisher
1	ust.hk	university
1	uta.edu	university
1	u-tokyo.ac.jp	university
3	utoronto.ca	university
1	uu.nl	university
1	uv.es	university
1	uwaterloo.ca	university
1	uwo.ca	university
1	viktoria.se	business
1	vliz.be	publisher
1	vu.nl	university
1	washington.edu	university
1	webscience.org	university
2	wisc.edu	university
1	wormbook.org	publisher
1	wustl.edu	university
1	yale.edu	university
1	ym.edu.tw	university
1	yorku.ca	university

APPENDIX C: 2012 Source Classification

2012 sources providing full text to articles with counts and classification of sources

Count	***Source***	***Type***
1	128.131.166.46	university
1	141.115.28.2	university
1	141.213.232.243	university
1	144.122.146.136	university
1	144.206.159.178	business
1	203.189.120.190	national
1	aacrjournals.org	publisher
1	adelaide.edu.au	university
1	adolphus.me.uk	business
1	adrprovita.com	business
1	ahajournals.org	publisher
1	ajcn.org	publisher
1	alphamedpress.org	publisher
2	ama-assn.org	publisher
4	ametsoc.org	publisher
2	anu.edu.au	university
1	apa.org	publisher
1	archives-ouvertes.fr	national
1	archybrid.com	personal
1	arizona.edu	university
26	arxiv.org	university
1	aspetjournals.org	publisher
1	asu.edu	university
1	au.dk	university
1	aut.ac.ir	university
1	baksheev.com.ua	personal
1	benthamscience.com	publisher
2	biologists.org	publisher
11	biomedcentral.com	publisher
1	birdflumanual.com	business
1	bris.ac.uk	university
1	brown.edu	university
1	bu.edu	university
1	cam.ac.uk	university
1	cancer.gov	government
1	cas.cz	university
1	chronobiology.ch	university
1	cicese.mx	university
2	cjb.net	unknown

Count	***Source***	***Type***
1	cmu.edu	university
1	computer.org	publisher
1	corgentum.com	business
1	cornell.edu	university
1	craigmcclain.com	personal
1	cshlp.org	business
1	cw.com.tw	publisher
1	cyganiak.de	personal
1	dbkgroup.org	university
1	dicp.ac.cn	university
1	digaden.edu.mx	university
3	dtic.mil	government
1	dtu.dk	university
2	duke.edu	university
1	dundee.ac.uk	university
1	dur.ac.uk	university
1	econmsu.org	university
1	emis.de	publisher
1	emory.edu	university
2	epfl.ch	university
1	fahamu.org	activism
1	fgv.br	university
1	fhv.at	university
1	francetelecom.fr	unknown
1	gatech.edu	university
1	gel.org.br	university
1	genego.com	publisher
3	genetics.org	publisher
1	gmu.edu	university
5	harvard.edu	university
1	hawaii.edu	university
1	heatherlench.com	personal
2	hematologylibrary.org	publisher
1	hindawi.com	publisher
1	iadis.net	publisher
1	ias.ac.in	publisher
1	ic.ac.uk	university
1	idei.fr	university
1	Idep-fr.org	university
1	igem.org	organization
1	iiarjournals.org	publisher
2	illinois.edu	university
1	imarpe.pe	university

Count	***Source***	***Type***
1	infn.it	university
1	intmedpress.com	publisher
1	isciii.es	national
1	ismni.org	publisher
1	isprs.org	publisher
1	itcj.edu.mx	university
1	ithaca.edu	university
1	iub.edu	university
3	jbc.org	publisher
1	jbjs.org	publisher
1	jee.org	publisher
2	jhu.edu	university
1	jmu.edu	university
2	jneurosci.org	publisher
1	joplink.net	publisher
1	ktu.edu	university
1	latech.edu	university
1	le.ac.uk	university
1	letunic.com	personal
1	liu.se	university
1	loria.fr	university
1	lth.se	university
1	lyellcollection.org	publisher
1	mahidol.ac.th	university
1	manchester.ac.uk	university
1	marquette.edu	university
1	mcponline.org	publisher
1	miami.edu	university
3	mit.edu	university
1	mpg.de	university
1	mshri.on.ca	university
1	mvr1.com	personal
1	nanofemtolab.qc.ca	university
1	nber.org	business
1	nb.rs	national
1	ncku.edu.tw	university
1	ncsu.edu	university
1	neuromorphs.net	publisher
40	nih.gov	government
2	nips.cc	publisher
1	Nslij-genetics.org	unknown
1	ntnu.no	university
1	ntu.edu.tw	university

Count	***Source***	***Type***
2	nyu.edu	university
1	otago.ac.nz	university
1	ou.edu	university
1	ovgu.de	university
12	oxfordjournals.org	publisher
1	pacomlan.com	unknown
1	pasteur.ac.ir	unknown
1	perceptionandaction.com	university
1	persoenlichkeitspsychologie-potsdam.de	university
3	petra.ac.id	university
2	physiology.org	publisher
1	physoc.org	publisher
1	pku.edu.cn	university
5	plos.org	publisher
11	pnas.org	publisher
1	pnexpert.com	personal
1	polytechnique.fr	university
1	proberts.net	personal
5	psu.edu	university
1	psykiatriskforskning.dk	university
1	Psy-net.be	university
1	qualcomm.de	unknown
1	rachel.org	business
1	rclis.org	university
1	rhbnc.ac.uk	university
1	rockefeller.edu	university
1	rockymedia.net	unknown
3	royalsocietypublishing.org	publisher
1	rsna.org	publisher
1	rug.nl	university
1	Ruhr-uni-bochum.de	university
1	ru.nl	university
4	rupress.org	publisher
1	rutgers.edu	university
2	sagepub.com	business
2	santafe.edu	university
2	scielosp.org	national
2	sgmjournals.org	publisher
1	shouxi.net	unknown
1	sigcomm.org	publisher
1	slu.se	university
1	soton.ac.uk	university
2	stanford.edu	university

Count	***Source***	***Type***
1	sunysb.edu	university
1	swarthmore.edu	university
1	syr.edu	university
1	technion.ac.il	university
1	toronto.edu	university
1	tribler.org	business
1	tuc.gr	university
1	tue.nl	university
1	tugraz.at	university
1	tum.de	university
1	ubc.ca	university
1	ucalgary.ca	university
1	ucl.ac.uk	university
2	ucla.edu	university
1	uconn.edu	university
1	ucsc.edu	university
2	ucsd.edu	university
1	ucsf.edu	university
1	uea.ac.uk	university
2	ugr.es	university
1	umassmed.edu	university
1	umd.edu	university
2	umich.edu	university
1	unam.mx	university
1	uned.es	university
1	unicamp.br	university
1	Uni-dortmund.de	university
1	Uni-goettingen.de	university
1	Uni-kl.de	university
1	uni-muenchen.de	university
1	Univ-bpclermont.fr	university
1	Univ-lyon1.fr	university
1	unl.edu	university
1	uoa.gr	university
1	uoregon.edu	university
1	uottawa.ca	university
2	upenn.edu	university
1	uq.edu.au	university
1	usc.edu	university
2	usda.gov	government
1	usenix.org	publisher
1	usp.br	university
1	ust.hk	university

Count	***Source***	***Type***
1	usu.edu.ru	university
2	uta.edu	university
1	utexas.edu	university
1	utk.edu	university
1	u-tokyo.ac.jp	university
3	utoronto.ca	university
1	utwente.nl	university
1	uu.nl	university
1	uv.es	university
1	uwo.ca	university
1	uzh.ch	university
1	vliz.be	publisher
1	vt.edu	university
2	washington.edu	university
1	whoi.edu	university
1	wormbook.org	publisher
1	yale.edu	university
1	yorku.ca	university

APPENDIX D: Predicted Probabilities

The following calculations are summarized in Table 13.

Using the logistic equation for the 2011 article full text model (Table 11),

$$\begin{gathered} Logit(p) = 120.2 + 0.0887(Authors) + 0.0425(Pub\,Year) \\ -0.1023(Post\,Year) + 0.0015(Citations\,2010) \end{gathered} \tag{8}$$

we can find the probability that an article is found full text for the first quartile, the median, and the third quartile values of the predictor variables.

For the first quartile values:

$$\begin{gathered} Predicted\,logit(p) = 120.2 + 0.0887(2) + 0.0425(1998) \\ -0.1023(2007) + 0.0015(4.75) = -0.0166 \end{gathered} \tag{9}$$

And then the predicted probability is:

$$Predicted\ probability = \frac{e^{(-0.0166)}}{\left(1+e^{(-0.0166)}\right)} = 0.4959 = 49.59\,\% \tag{10}$$

For the median values:

Three authors, was published in 2004, was posted to CiteULike in 2008, and has a citation count of 23:

$$\begin{gathered} Predicted\,logit(p) = 120.2 + 0.0887(3) + 0.0425(2004) \\ -0.1023(2008) + 0.0015(23) = 0.2522 \end{gathered} \tag{11}$$

And then the predicted probability is:

$$Predicted\ probability = \frac{e^{(0.2522)}}{\left(1+e^{(0.2522)}\right)} = 0.5627 = 56.27\,\% \tag{12}$$

For the third quartile values:

$$\begin{gathered}\textit{Predicted logit}\,(p)=120.2+0.0887\,(4)+0.0425\,(2007)\\-0.1023\,(2009)+0.0015\,(72)=0.1492\end{gathered}\tag{13}$$

And then the predicted probability is:

$$\textit{Predicted probability}=\frac{e^{(0.4396)}}{\left(1+e^{(0.4396)}\right)}=0.6082=60.82\%\tag{14}$$

Likewise, we can repeat this for the 2012 article full text model (Table 12), where the logistic regression equation is:

$$\begin{gathered}\textit{Logit}\,(p)=88.4637+0.0100\,(\textit{Authors})+0.0473\,(\textit{Pub Year})\\-0.0911\,(\textit{Post Year})+0.0016\,(\textit{Citations}\,2011)\end{gathered}\tag{15}$$

For the first quartile values:

$$\begin{gathered}\textit{Logit}\,(p)=88.4637+0.0100\,(2)+0.0473\,(1998)\\-0.0911\,(2007)+0.0016\,(7.0)=0.1626\end{gathered}\tag{16}$$

$$\textit{Predicted probability}=\frac{e^{(0.1626)}}{\left(1+e^{(0.1626)}\right)}=0.5406=54.06\%\tag{17}$$

For the median values:

$$\begin{gathered}\textit{Logit}\,(p)=88.4637+0.0100\,(3)+0.0473\,(2004)\\-0.0911\,(2008)+0.0016\,(28)=0.3989\end{gathered}\tag{18}$$

$$Predicted\ probability = \frac{e^{(0.3989)}}{\left(1+e^{(0.3989)}\right)} = 0.5984 = 59.84\,\% \tag{19}$$

For the third quartile values:

$$\begin{gathered} Logit(p) = 88.4637 + 0.0100(4) + 0.0473(2007) \\ -0.0911(2009) + 0.0016(82.5) = 0.5469 \end{gathered} \tag{20}$$

$$Predicted\ probability = \frac{e^{(0.5469)}}{\left(1+e^{(0.5469)}\right)} = 0.6334 = 63.34\,\% \tag{21}$$

BIBLIOGRAPHY

Abbott, A. (2004). *Methods of discovery: Heuristics for the social sciences.* Contemporary Societies. New York: W. W. Norton and Company.

ACRL. (2000). *Information literacy competency standards for higher education.* Retrieved from http://www.ala.org/acrl/standards/informationliteracycompetency

Adkins, D., & Bala, E. (2004). Public library outreach as a function of staffing and metropolitan location. *Library and Information Science Research, 26*, 338-350. doi:10.1016/j.lisr.2004.01.001

Aguillo, I. F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics*, 91, 343-351. doi:10.1007/s11192-011-0582-8

Akeroyd, J. (2001). The future of academic libraries. *Aslib Proceedings*, *53*(3), 79-84. doi:10.1108/EUM0000000007041

Albert, K. M. (2006). Open access: Implications for scholarly publishing and medical libraries. *Journal of the Medical Library Association*, *94*(3), 253-262. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525322/

Alder, H. L., & Roessler, E. B. (1968). *Introduction to probability and statistics* (4th ed.). San Francisco: W.H. Freeman and Company.

ARL Statistics: 2004-05: A compilation of statistics from the one hundred and twenty-three members of the Association of Research Libraries. (2006). Washington, D.C.: Association of Research Libraries. Retrieved from http://www.arl.org/

ARL Statistics: 2010-2011. (2012). Washington, D.C.: Association of Research Libraries. Retrieved from the Association of Research Libraries web site:

http://www.arl.org/

Arlitsch, K. and O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratio of IRs in Google Scholar. *Library Hi Tech, 30*(1), 60-81. doi:10.1108/07378831211213210

Bailey, C. W. (2007). Open access and libraries. *Collection Management, 32,* 351-383. doi:10.1300/J105v32n03_07

Baldwin, V. A. (2009). Using Google Scholar to search for online availability of a cited article in engineering disciplines. *Issues in Science and Technology Librarianship, 56*. doi:10.5062/F4WM1BBC

Bates, M. (1981). Search techniques. *Annual Review of Information Science and Technology, 16*, 139-169.

Bergstrom, C. T. and Bergstrom, T. C. (2006). The economics of ecology journals. *Frontiers in Ecology and the Environment, 4*(9), 488-495. doi:10.1890/1540-9295(2006)4[488:TEOEJ]2.0.CO;2

Binmore, K. (1994). *Game theory and the social contract: Playing fair* (Vol. 1). Cambridge, MA: MIT Press.

Binmore, K. (2007). *Game theory: A very short introduction*. Oxford: Oxford University Press.

Binmore, K. (2008). *Rational Decisions.* Princeton: Princeton University Press.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216-1227. doi:10.1002/asi.20077

Black, A. (2007). Mechanization in libraries and information retrieval: Punched cards and microfilm before the widespread adoption of computer technology in libraries. *Library History, 23*, 291-299. doi:10.1179/174581607x254785

Bogers, T., & van den Bosch, A. (2008). Recommending scientific articles using CiteULike. In *Proceedings of the 2008 ACM conference on Recommender systems*, (pp. 287–290). New York, NY: ACM.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*, 2-72. doi:10.1002/aris.1440360102

Bornmann, L., & Hans-Dieter, D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80. doi:10.1108/00220410810844150

Borrego, A., & Fry, J. (2012). Measuring researchers' use of scholarly information through social bookmarking data: A case study of BibSonomy. *Journal of Information Science, 38*(3), 297-308. doi:10.1177/0165551512438353

Bosch, S. and Henderson, K. (2012, April 30). Coping with the terrible twins: Periodicals price survey 2012. *Library Journal*. Retrieved from http://lj.libraryjournal.com/2012/04/funding/coping-with-the-terrible-twins-periodicals-price-survey-2012/

Boslaugh, S. (2012). *Statistics in a Nutshell (2nd ed.)*. Beijing: O'Reilly.

Brandstätter, E., & Brandstätter, H. (1996). What's money worth? Determinants of the subjective value of money. *Journal of Economic Psychology, 17*(4), 443-464.

doi:10.1016/0167-4870(96)00019-0

Broadus, R. N. (1987). Toward a definition of "bibliometrics". *Scientometrics, 12*(5-6), 373-379. doi:10.1007/BF02016680

Brookes, B. C. (1969). Bradford's Law and the bibliography of science. *Nature, 224*(5223), 953-956. doi: 10.1038/224953a0

Budd, J. (1986). Characteristics of written scholarship in American literature: A citation study. *Library and Information Science Research, 8*(2), 189-211.

Budd, J. M. (1992). Bibliometrics: A method for the study of the literature of higher education. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 345-378). New York: Agathon Press.

Budd, J. M. (2002). Serials prices and subscriptions in the social sciences. *Journal of Scholarly Publishing, 33*(2), 90-101. doi:10.3138/jsp.33.2.90

Budd, J. M. (2009). Academic library data from the United States: An examination of trends. *Libres: Library and Information Science Research Electronic Journal, 19*(2), 1-21.

Budd, J. M. (2012). Scholarly communication's mess: Can economic analysis help? *Libres: Library and Information Science Research Electronic Journal, 22*(1), 1-17.

Burns, C. S., Lana, A., & Budd, J. M. (2013). Institutional repositories: Exploration of costs and value. *D-Lib Magazine, 19*(1/2). doi:10.1045/january2013-burns

Bush, V. (1945, July). As we may think. *Atlantic Monthly.* Retrieved from http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

Capocci, A., & Caldarelli, G. (2008). Folksonomies and clustering in the collaborative system CiteULike. *Journal of Physics A: Mathematical and Theoretical, 41*. doi:10.1088/1751-8113/41/22/224016

Carpenter, K. E. (1996). A library historian looks at librarianship. *Daedalus, 125*(4), 77-102.

Carrigan, D. P. (1995). Toward a theory of collection development. *Library Acquisitions: Practice & Theory, 19*(1):97-106. doi:10.1016/0364-6408(94)00056-2

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635-645.

Cave, E. M. (2005). A normative interpretation of expected utility theory. *The Journal of Value Inquiry, 39*(3), 431-441. doi:10.1007/s10790-006-7525-2

Chacko, G. K. (1991). *Decision-making under uncertainty: An applied statistics approach.* New York: Praeger.

Chan, L. (2004). Supporting and enhancing scholarship in the digital age: The role of open-access institutional repositories. *Canadian Journal of Communication, 29*(3). Retrieved from http://www.cjc-online.ca/index.php/journal/article/view/1455/1579

Chen, X. (2010). Google Scholar's dramatic coverage improvement five years after debut. *Serials Review, 36*(4), 221-226. doi:10.1016/j.serrev.2010.08.002

CiteULike. (2010a). *Available datasets* [Accessing CiteULike datasets]. Retrieved from http://www.citeulike.org/faq/data.adp.

CiteULike. (2010b). *Everyone's library* [Listing of recent entries]. Retrieved from http://www.citeulike.org/all/page/1.

CiteULike. (2010c). How to post a paper to CiteULike [Instructions for browser bookmarklet]. Retrieved from http://www.citeulike.org/post.

Collins, C. S. and Walters, W. H. (2010). Open access journals in college library collections. *The Serials Librarian, 59*(2), 194-214. doi:10.1080/03615261003623187

Corrado, E. M. (2005). The importance of open access, open source, and open standards for libraries. *Issues in Science and Technology, 42*(Spring 2005). Retrieved from http://istl.org/05-spring/article2.html

Cothran, T. (2011). Google scholar acceptance and use among graduate students: A quantitative study. *Library & Information Science Research, 33*(4), 293-301. doi:10.1016/j.lisr.2011.02.001

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

Cronin, B. (2001). Bibliometrics and beyond: Some thoughts on web-based citation analysis. *Journal of Information Science, 27*, 1-7. doi:10.1177/016555150102700101

Cronin, B., Shaw, D., & La Barre, K. L. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society* *for Information Science and Technology, 54*(9), 855-871. doi:10.1002/asi.10278

Cronin, B, & Franks, S. (2006). Trading cultures: Resource mobilization and service rendering in the life sciences as revealed in the journal article's paratext. *Journal of the American Society for Information Science and Technology, 57*(14), 1909-1918. doi:10.1002/asi.20407.

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: Randomized controlled trial. *BMJ, 337*(a568). doi:10.1136/bmj.a568

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics*. Lanham: Scarecrow Press.

De Bruin, B. (2005). Game theory in philosophy. *Topoi*, *24*(2), 197-208. doi:10.1007/s11245-005-5055-3

Ding, Y., & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing & Management*, *47*(1), 80-96. doi:10.1016/j.ipm.2010.01.002

Dixit, A. K. and Skeath, S. (2004). *Games of strategy*. New York: W. W. Norton & Company.

Dourish, P. (2001). *Where the action is: The foundations of embodied interaction.* Cambridge: MIT Press.

Dowd, S. T. (1990). Library cooperation: Methods, models to aid information access. *Journal of Library Administration, 12*(3), 63-81.

Drott, M. C. (2006). Open access. *Annual Review of Information Science and*

Technology, 40, 79-109. doi:10.1002/aris.1440400110

Dufwenberg, M. (2010). Psychological games. In S. N. Durlauf & L. E. Blume (Eds.), *Game theory* (pp. 272-278). New York: Palgrave MacMillan.

Eisenstein, E. (1983). *The printing revolution in early modern Europe*. New York: Cambridge University Press.

Ellis, D. (1989). A behavioral approach to information retrieval system design. *Journal of Documentation, 45*(3), 171-212.

Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation, 49*(4), 356-369.

Emamy, K., & Cameron, R. (2007). Citeulike: A researcher's social bookmarking service. *Ariadne, 51.* Retrieved from http://www.ariadne.ac.uk/issue51/emamy-cameron.

Enger, K. B. (2009). Using citation analysis to develop core book collections in academic libraries. *Library & Information Science Research, 31*(2), 107-112. doi:10.1016/j.lisr.2008.12.003

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology, 4*(5), e157. doi:10.1371/journal.pbio.0040157

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338-342. doi:10.1096/fj.07-9492LSF

Farber, S. (1998). Undesirable facilities and property values: A summary of empirical studies. *Ecological Economics, 24*(1), 1-14. doi:10.1016/S0921-8009(97)00038-4

Fidczuk, R., Beebe, L., & Wallas, P. (2007). Today's journal cost: Print vs. online. *Serials Librarian, 52*(3/4), 341-348. doi:10.1300/J123v52n03_15

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* Los Angeles: Sage.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108-111. doi:10.1126/science.122.3159.108

Gargouri, Y., Hajjem, C., Lariviere, V., Gingras, Y, Carr, Y, Carr, L., Brody, T., & Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE, 5*(10), e13636. doi:10.1371/journal.pone.0013636

Google. (n.d.). Library support. Retrieved from http://scholar.google.com/intl/en-US/scholar/libraries.html

Greco, A. N., Wharton, R. M., Estelami, H., & Jones, R. F. (2006). The state of scholarly publishing: 1981-2000. *Journal of Scholarly Publishing, 37*(3), 155-214.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software, 40*(3), 1-25. Retrieved from http://www.jstatsoft.org/v40/i03/

Grüne-Yanoff, T., & Schweinzer, P. (2008). The roles of stories in applying game theory. *Journal of Economic Methodology, 15*(2), 131-146. doi:10.1080/13501780802115075

Hamlin, A. T. (1981). *The university library in the United States*. Philadelphia: University of Pennsylvania Press.

Harloe, B., & Budd, J. M. (1994). Collection development and scholarly communication in the era of electronic access. *The Journal of Academic Librarianship, 20*(2), 83-87. doi:10.1016/0099-1333(94)90043-4

Harnad, S., Brody, T., Valliéres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C., & Hilf, E. R. (2008). The access/impact problem and the green and gold roads to open access: An update. *Serials Review, 34*(1), 36-40. http://dx.doi.org/10.1016/j.serrev.2007.12.005

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Harzing, A. W. K., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics, 8*, 61-73. doi:10.3354/esep00076

Hausman, D. M. (2000). Revealed preference, belief, and game theory. *Economics and Philosophy, 16*(1), 99-115. doi:10.1017/S0266267100000158

Hausman, D. M. (2005). Sympathy, commitment, and preference. *Economics and philosophy, 21*(1), 33-50. doi:10.1017/S0266267104000379

Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics, 5*(3), 446-457. doi:10.1016/j.joi.2011.04.002

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology, 61*(2), 310-318. doi:10.1002/asi.21256

Herrera, G. (2011). Google scholar users and user behaviors: An exploratory study. *College & Research Libraries, 72*(4), 316-330.

Hill, H. (2009). *Outsourcing the public library: A critical discourse analysis*. (Unpublished doctoral dissertation). University of Missouri. Retrieved from http://hdl.handle.net/10355/6126

Howland, J. L., Wright, T. C., Boughan, R. A., & Roberts, B. C. (2009). How scholarly is Google Scholar? Comparison to library databases. *College & Research Libraries, 70*(3), 227–234.

Hull, D., Pettifer, S. R., & Kell, D. B. (2008). Defrosting the digital library: Bibliographic tools for the next generation web. *PLoS Computational Biology, 4*(10), e1000204. doi:10.1371/journal.pcbi.1000204.

Julien, H., & Genuis, S. K. (2011). Librarians' experience of the teaching role: A national survey of librarians. *Library and Information Science Research, 33*(2), 103-111. doi:10.1016/j.lisr.2010.09.005

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292. doi:10.2307/1914185

Kilgour, F. G. (1939). A new punched card for circulation records. *Library Journal, 64*(4), 131-133.

King, J. E. (2008). Binary logistic regression. In J. Osborne (Ed.) *Best practices in quantitative methods* (pp. 358-384). Thousand Oaks, CA: Sage Publications.

Kipp, M. E. I. (2011). User, author and professional indexing in context: An exploration of tagging practices on CiteULike. *Canadian Journal of Information and Library Science, 35*(1), 17-48. doi:10.1353/ils.2011.0008

Kirkwood, H. P. & Kirkwood, M. C. (2011). Historical research: Historical abstracts with full text or Google Scholar. *Online: Exploring Technology & Resources for Informational Professionals, 35*(4), 28-32.

Kling, R., & Callahan, E. (2003). Electronic journals, the Internet, and scholarly communication. *Annual Review of Information Science and Technology, 37*, 127-177. doi:10.1002/aris.1440370105

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58*(7), 1055-1065. doi:10.1002/asi.20584

Kress, N., Bosque, D. D., & Ipri, T. (2011). User failure to find known library items. *New Library World, 112*(3/4), 150-170. doi:10.1108/03074801111117050

Kruschke, J. K. (2011). *Doing Bayesian analysis: A tutorial with R and BUGS*. Amsterdam: Academic Press.

Lancaster, F. W. (1978). *Toward paperless information systems*. New York: Academic Press.

Lewis, D. W. (2012). The inevitability of open access. *College and Research Libraries, 73*(5), 493-506.

Licklider, J. C. R. (1965). *The library of the future*. Cambridge, MA: MIT Press.

Lindley, D. V. (1971). *Making decisions.* London: Wiley.

Line, M. B. (1970). The 'half-life' of periodical literature: Apparent and real obsolescence. *Journal of Documentation, 26*(1), 46-54.

Lingel, J. (2012). Occupy Wall Street and the myth of technological death of the library. *First Monday, 17*(8). doi:10.5210/fm.v17i8.3845

Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *portal: Libraries and the Academy, 3*(2), 327-336. doi:10.1353/pla.2003.0039

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology, 61*(1), 1-13. doi:10.1002/asi.21228

McCain, K. W., & Bobick, J. E. (1981). Patterns of journal use in a departmental library: A citation analysis. *Journal of the American Society for Information Science, 32*(4), 257-267.

McGuigan, G. S., & Russell, R. D. (2008). The business of academic publishing: A strategic analysis of the academic journal publishing industry and its impact on the future of scholarly publishing. *Electronic Journal of Academic and Special Librarianship, 9*(3). Retrieved from http://southernlibrarianship.icaap.org/content/v09n03/mcguigan_g01.html

Meho, L. I. (2007). The rise and rise of citation analysis. *Physics World, 20*, 32-36.

Meho, L. I., & Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American Society for Information Science and Technology, 54*(6), 570-587. doi:10.1002/asi.10244

Meho, L. I. & Yang, K. (2007). Impact of data sources on citation counts and rankings of lis faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105–2125. doi:10.1002/asi.20677

Michalak, S. C. (2012). This changes everything: Transforming the academic library. *Journal of Library Administration, 52*(5):411-423. doi:10.1080/01930826.2012.700801

Mitchell, B. A. (2007). Boston Library catalogues, 1850-1875. In T. Augst, & K. Carpenter (Eds.), *Institutions of reading: The social life of libraries in the United States* (pp. 119-147). Amherst: University of Massachusetts Press.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*(9), 810-832.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, *10*(3), 303-320.

Moran, B. B. (2001). Restructuring the university library: A North American perspective. *Journal of Documentation, 57*(1), 100-114. doi:10.1108/EUM0000000007079

Moulaison, H. L, & Burns, C. S. (2012). Organization or conversation in Twitter: A case study of chatterboxing. *Proceedings of the American Society for Information Science and Technology, 49*(1), 1-11. doi:10.1002/meet.14504901185

Mullen, L. B., & Hartman, K. A. (2006). Google Scholar and the library web site: The early response by ARL libraries. *College and Research Libraries, 67*(2), 106-122.

Narin, F., & Moll, J. K. (1977). Bibliometrics. *Annual Review of Information Science and Technology, 12*, 35-58.

National Institutes of Health. (n.d.). National Institutes of Health public access. Policy overview. Retrieved from http://publicaccess.nih.gov/.

Neuhaus, C., Neuhaus, E., & Asher, A. (2008). Google Scholar goes to school: The presence of Google Scholar and university web sites. *The Journal of Academic Librarianship, 34*(1), 39-51. doi:10.1016/j.acalib.2007.11.009

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17*, 857-872.

Nickel, P. J. (2009). Trust, staking, and expectations. *Journal for the Theory of Social Behaviour, 39*(3), 345-362. doi:10.1111/j.1468-5914.2009.00407.x

Nicolaisen, J. (2002). The J-shaped distribution of citedness. *Journal of Documentation, 58*(4), 383-395. doi:10.1108/00220410210431118

Nicolaisen, J. (2003). The social act of citing: Towards new horizons in citation theory. *Proceedings of the American Society for Information Science and Technology, 40*(1), 12-20. doi:10.1002/meet.1450400102

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology, 41*, 609-641. doi:10.1002/aris.2007.1440410120

Niu, X., & Hemminger, B. M. (2012). A study of factors that affect the information-seeking-behavior of academic scientists. *Journal of the American Society for Information Science and Technology, 63*(2), 336-353. doi:10.1002/asi.21669

Niu, X., Hemminger, B. M., Lown, C., Adams, S., Brown, C., Level, A., McLure, M., Powers, A., Tennant, M. R., & Cataldo, T. (2010). National study of information seeking behavior of academic research in the United States. *Journal of the American Society for Information Science and Technology, 61*(5), 869-890. doi:10.1002/asi.21307

Norris, M., Oppenheim, C., and Rowland, F. (2008). Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR. *Online Information Review, 32*(6), 709-715. doi:10.1108/14684520810923881

Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *Libri, 55*, 170-180. doi:10.1515/LIBR.2005.170

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science, 289*, 1773-1775. doi:10.1126/science.289.5485.1773

Nyquist, C. (2010). An academic librarian's response to the "ITHAKA faculty survey 2009: Key strategic insights for libraries, publishers, and societies". *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve, 20*(4), 275-280. doi:10.1080/1072303X.2010.508419

Oberg, L. R., Mentges, M. E., McDermott, P. N., & Harusadangkul, V. (1992). The role, status, and working conditions of paraprofessionals: A national survey of academic libraries. *College and Research Libraries*, *53*(3), 215-238.

O’Reilly, T. (2005). What is web 2.0: Design patterns and business models for the next generation of software. Retrieved from http://oreilly.com/pub/a/web2/archive/what-is-web-20.html

Parker, R. H. (1936). The punched card method in circulation work. *The Library*

Journal, 61, 903-905.

Peek, R. (2012). Digital Public Library of America. *Information Today, 29*(2), 24.

Phillips, L. D. (1973). *Bayesian statistics for social scientists*. London: Nelson.

Pinfield, S. (2005). A mandate to self archive? The role of open access institutional repositories. S*erials, 18*(1), 30-34. doi:10.1629/1830

Plutchak, T. S. (2012). Breaking the barriers of time and space: the dawning of the great age of librarians. *Journal of the Medical Library Association, 100*(1), 10-19. doi:10.3163/1536-5050.100.1.004

Pomerantz, J. (2006). Google Scholar and 100% availability of information. *Information Technology and Libraries, 25*(1), 52-56.

Pomerantz, J., & Marchionini, G. (2007). The digital library as place. *Journal of Documentation, 63*(4), 505-533. doi:10.1108/00220410710758995

Price, D. J. D. S. (1986). *Little science, big science ...and beyond*. New York: Columbia University Press.

Priem, J. (2013). Scholarship: Beyond the paper. *Nature, 495*, 437-440. doi:10.1038/495437a

Priem, J., & Hemminger, B. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social web. *First Monday, 15*(7). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/2874/2570

Romero, L. (2008). Confirming suspicions: An analysis of original communication studies journal price data. *Collection Management, 33*(3), 189-218. doi:10.1080/01462670802045525

R Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica, 68*(5), 1281-1292. doi:10.1111/1468-0262.00158

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty.* Reading, MA: Addison-Wesley.

Ritzberger, K. (2002). *Foundations of non-cooperative game theory*. Oxford University Press.

Rosenthal, E. C. (2011). *The complete idiot's guide to game theory*. New York: Alpha Books.

Ross, C. S. (2009). Reader on top: Public libraries, pleasure reading, and models of reading. *Library Trends, 57*(4), 632-656. doi:10.1353/lib.0.0059

Ross, D. (2010). Game theory. *The Stanford Encyclopedia of Philosophy (Winter 2012 Ed.)*. Retrieved from http://plato.stanford.edu/archives/win2012/entries/game-theory/

Sapp, G., & Gilmour, R. (2002). A brief history of the future of academic libraries: Predictions and speculations from the literature of the profession, 1975 to 2000-part one, 1975 to 1989. *portal: Libraries and the Academy, 2*(4):553-576. doi:10.1353/pla.2002.0086

Sapp, G., & Gilmour, R. (2003). A brief history of the future of academic libraries: Predictions and speculations from the literature of the profession, 1975 to 2000-part two, 1990 to 2000. *portal: Libraries and the Academy, 3*(1):13-34.

doi:10.1353/pla.2003.0008

Savolainen, R. (2012). Expectancy-value beliefs and information needs as motivators for task-based information seeking. *Journal of Documentation, 68*(4), 492-511. doi:10.1108/00220411211239075

Schonfeld, R. C., & Housewright, R. (2010). Faculty survey 2009: Key strategic insights for libraries, publishers, and societies. Retrieved from http://www.sr.ithaka.org/research-publications/faculty-survey-2009

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica, 57*(3), 571-587.

Sen, A. (1973). Behaviour and the concept of preference. *Economica, 40*(159), 241-259.

Sen, A. (1977). Rational fools: A critique of the behavioural foundations of economic theory. *Philosophy and Public Affairs, 6*(4), 317-344.

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization*. New York: Doubleday.

Sennyey, P., Ross, L., & Mills, C. (2009). Exploring the future of academic libraries. *The Journal of Academic Librarianship, 35*(3):252-259. doi:10.1016/j.acalib.2009.03.003

Shiflett, O. L. (1981). *Origins of American academic librarianship*. Norwood, NJ: Ablex Publishing Corporation.

Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*(1), 99-118.

Simon, H. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1-

19.

Sin, S-C. J. (2011). International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980-2008. *Journal of the American Society for Information Science and Technology, 62*(9), 1770-1783. doi:10.1002/asi.21572

Sin, S-C. J., & Kim, K-S. (2008). Use and non-use of public libraries in the information age: A logistic regression analysis of household characteristics and library services variables. *Library and Information Science Research, 30*, 207-215. doi:10.1016/j.lisr.2007.11.008

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327-340. doi:10.1177/030631277800800305

Smith, E. (1990). *The librarian, the scholar, and the future of the research library.* New York: Greenwood Press.

Smith, L. C. (1981). Citation analysis. *Library Trends*, *30*, 83-106.

Suber, P. (2004, December 29). A very brief introduction to open access. Retrieved from http://www.earlham.edu/~peters/fos/brief.htm

tbogers (2009). Science papers that interest you. Retrieved December 01, 2009, from http://blog.citeulike.org/?p=11

Tenopir, C., King, D. W., Spencer, J., and Wu, L. (2009). Variations in article seeking and reading patterns of academics: What makes a difference? *Library & Information Science Research*, *31*(3), 139-148. doi:10.1016/j.lisr.2009.02.002

Thelwall, M., & Harries, G. (2004). Do the web sites of higher rated scholars have significantly more online impact? *Journal of the American Society for*

Information Science and Technology, *55*(2), 149-159. doi:10.1002/asi.10362

Theng, Y-L., & Sin, S-C. J. (2012). Analysing the effects of individual characteristics and self-efficacy on users' preferences for system features in relevance judgment. *Information Research, 17*(4). Retrieved from http://informationr.net/ir/17-4/paper536.html#.UQr4qVn1SJg

Tibbo, H. R., Clemens, R., & Hank, C. (2009). Introduction. *Bulletin of the American Society for Information Science and Technology, 35*(4), 11-31.

Tenopir, C. (2012). Beyond usage: Measuring library outcomes and value. *Library Management, 33*(1/2), 5-13. doi:10.1108/01435121211203275

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124

Vaughan, L. (2001). *Statistical methods for the information professional: A practical, painless approach to understanding, using, and interpreting statistics*. Medford, NJ: Information Today.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: What is the difference? *Journal of the American Society for Information Science and Technology, 54*(14), 1312-1322. doi:10.1002/asi.10338

Vaughan, L., & Shaw, D. (2008). A new look at evidence of scholarly citation in citation indexes and from web sources. *Scientometrics, 74*(2), 317-330. doi:10.1007/s11192-008-0220-2

Vieira, E. S., & Gomes, J. A. N. F. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics, 4,* 1-13.

doi:10.1016/j.joi.2009.06.002.

Wainer, J., Oliveira, H., & Anido, R. (2011). Patterns of bibliographic references in the ACM published papers. *Information Processing & Management, 47*(1), 135-142. doi:10.1016/j.ipm.2010.07.002

Walters, W. H. (2009). Google Scholar search performance: Comparative recall and precision. *portal: Libraries and the Academy, 9*(1), 5–24. doi:10.1353/pla.0.0034

Walters, W. H. & Linvill, A. C. (2011a). Bibliographic index coverage of open-access journals in six subject areas. *Journal of the American Society for Information Science and Technology, 62*(8), 1614-1628. doi:10.1002/asi.21569

Walters, W. H., & Linvill, A. C. (2011b). Characteristics of open access journals in six subject areas. *College and Research Libraries, 72*(4), 372-392.

White, H. D. (2007a). Combining bibliometrics, information retrieval, and relevance theory, part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology, 58*(4), 536-559. doi:10.1002/asi.20543

White, H. D. (2007b). Combining bibliometrics, information retrieval, and relevance theory, part 2: Some implications for information science. *Journal of the American Society for Information Science and Technology, 58*(4), 583-605. doi:10.1002/asi.20542

White, H. D., & McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology*, 24, 119-186.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1-20. Retrieved on February 3, 2013 from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2009). *ggplot2: Elegant graphics for Data Analysis.* New York: Springer.

Wiegand, W. A. (1990). Research libraries, the ideology of reading, and scholarly communication, 1876-1900. In P. Dain, & J.Y. Cole (Eds.) *Libraries and scholarly communication in the United States: The historical dimension* (pp. 71-87). New York: Greenwood Press.

Wolfram, D. (2003). *Applied informetrics for information retrieval research.* Westport, Conn.: Libraries Unlimited.

Wouters, P. (1998). The signs of science. *Scientometrics*, 41, 225-241. doi:10.1007/BF02457980.

Yadamsuren, B., Paul, A., Wang, J., Wang, X., & Erdelez, S. (2008). Web ecology: Information needs of different user groups in the context of a community college website. *Proceedings of the American Society for Information Science and Technology, 45*(1), 1-4. doi:10.1002/meet.2008.14504503112

Yang, K., & Meho, L. I. (2006). Citation analysis: A comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology, 43*, 1-15. doi:10.1002/meet.14504301185

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley Press.

VITA

I was born in Springfield, Missouri in 1972 and my family moved to San Antonio, Texas in the summer of 1978. In 1991, I attended Monmouth College in Monmouth, Illinois, where I double majored in Philosophy & Religious Studies and English. After graduation, I worked as a cook and chef for twelve years. In 2007, I switched gears and pursued graduate study in library science at the University of Missouri. In my second year of graduate school, I realized that my interests favored teaching and research, and I decided to pursue a PhD in library and information science.

My research focuses on academic libraries and scholarly communication --- especially how the latter affects the former. I will continue to pursue this research in the fall of 2013 when I begin an appointment as an assistant professor at the University of Kentucky's School of Library and Information Science.