EXPLORING THE CONVIENIENCE VERSUS NECESSITY DEBATE REGARDING SCI-HUB USE IN THE UNITED STATES

by

John O. LaDue

AS, Monroe Community College, 1998

BA, Grove City College, 1999

MLIS, University of Pittsburgh, 2008

Submitted to the Graduate Faculty of
the School of Education in partial fulfillment
of the requirements for the degree of
Doctor of Education

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

John O. LaDue

It was defended on

May 16, 2018

and approved by

Dr. Lori Delale O'Conner, Assistant Professor, Center for Urban Education

Dr. Lindsay Page, Assistant Professor, Psychology in Education

Dr. Christinger Tomer, Associate Professor, School of Information Science

Dissertation Advisor: Dr. Linda DeAngelo, Associate Professor, Administrative and Policy Studies

EXPLORING THE CONVENIENCE VERSUS NECESSITY DEBATE REGARDING SCI-HUB USE IN THE UNITED STATES

John O. LaDue, Ed.D.

University of Pittsburgh, 2018

This study used multiple regression modeling to explore the relationship between Sci-Hub use in the United States and the characteristics of the areas surrounding the download requests. The purpose of this study was to examine Sci-Hub usage in the United States to explore the validity of academic journal publisher claims of convenience over necessity. This study was broken down into two parts: 1) how Sci-Hub download requests are related to the institutional characteristics of research-intensive universities and 2) how Sci-Hub download requests are related to the population of their geographic location. Convenience, for the purpose of this study, was based on Zipf's Principle of Least Effort.

In the first part of this study, universities were associated with Sci-Hub download requests within a 10-mile radius of the institution. The predictor variables for this section included an institution's journal expenditures, the size of the graduate student and faculty population, and the amount of research funding from NIH and NSF. Research funding was found to have a positive, significant relationship with Sci-Hub use when controlling for the other predictors. Additionally, an interaction between the amount of research funding and the size of graduate student and faculty population was included in the final model. Institutions with larger numbers of graduate students and faculty and higher levels of research funding were found to have the highest levels of Sci-Hub use. This interaction effect suggests that necessity may be more of a driver of Sci-Hub use than convenience.

In the second part of the study, Sci-Hub download requests were split up by the core-based statistical areas (CBSAs). The models in this section examined population size, the percentage of the population with an advanced degree, the number and type of higher education institutions, and the number of graduate students. Advanced degree holders were found to have a positive, significant relationship with Sci-Hub use when accounting for the other predictor variables. This finding may suggest necessity as people outside of higher education often do not have access to academic literature. Taken together, the two parts of the study suggest that necessity is likely driving people to use Sci-Hub.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First and foremost, I want to thank my advisor, Linda DeAngelo. You have not only been a good advisor, you have been a good friend. You have helped me through difficult times, both academically and personally. Without you dragging me across the finish line, I'm not sure I would have made it.

I want to thank my committee members: Lori Delale O'Connor, Lindsay Page, and Christinger Tomer. I appreciate all of the time and effort you have put forth for my benefit.

Thank you to Fran Yarger. You are not only the boss who gave me a shot and got me into higher education, you have been a mentor as my career has blossomed, and, most importantly, you have become one of my best friends. Without your advice and sympathetic ear through all the trials and tribulations of the past 10+ years, I'm not sure if I would have made it to this point.

I owe my parents, Roger and Margaret, a huge debt of gratitude. Without your love and encouragement, I likely would not have finished my Bachelor's degree, let alone be on the precipice of a Doctorate. Your support means more than you'll ever know.

Lastly, I want to thank my wife, Lindy, for bearing with me through multiple degrees. Without your patience and support, I would not have been able to manage a full-time job while going to school. I promise this is the last degree I will get.

1.0 INTRODUCTION

"Information is power. But like all power, there are those who want to keep it for themselves. The world's entire scientific and cultural heritage, published over centuries in books and journals, is increasingly being digitized and locked up by a handful of private corporations."

From Guerilla Open Access Manifesto, Swartz, 2008

In the greater context of higher education, academic literature is both of utmost importance and an afterthought. Through a macro lens, the role of academic literature is to disseminate and share knowledge and theory and help advance scientific pursuit. Through a personal, micro lens, academic literature is a primary factor in career advancement under the current publish or perish model of academia (Liebowitz, 2015). However, despite the systemic and personal value of academic literature, the industry that has arisen around it can be overlooked even by those who participate in it (Kocken & Wical, 2013).

As a major source of information in higher education, scholarly journals serve to disseminate the research findings and theories put forth by academics, but access to these journals is frequently limited to individuals and institutions willing and able to pay for it. Studies indicate that roughly three-quarters of academic literature reside behind paywalls (Khabsa & Giles, 2014; Piwowar et al., 2017). Paywalls are the name given artificial restrictions placed on materials to ensure revenue for accessing the content (Estok, 2011). For those without subscriptions or an affiliation with a subscribing institution, individual articles can cost upwards of $40, including an article I co-wrote (Saleh, Ratajeski, & Ladue, 2014) that would cost $42 for 24 hours of access. It was in this context that Alexandra Elbakyan, a graduate student in Kazakhstan at the time (Bohannon, 2016b), developed Sci-Hub in 2011. Sci-Hub is a repository that aims to host free copies of academic articles that can be used by academics, or anyone, to access research findings. What makes Sci-Hub controversial is that it circumvents copyright laws to offer free copies of materials that would otherwise be paywalled.

Publishers argue that Sci-Hub use, especially in highly developed nations, is a matter of convenience, not necessity (Bohannon, 2016b). They point to mechanisms such as site licenses, open access, and other means which will be discussed in further detail in this chapter, as legal, alternative methods to access academic literature (McNutt, 2016). However, these assertions of convenience over necessity are more theoretical than data-driven, as there have been limited studies that can provide insights into this phenomenon.

In this study, I explored Sci-Hub usage in the United States from two angles. First, the relationship between the number of Sci-Hub download requests near research institutions, the academic journal expenditures of that institution's library, and other institutional characteristics was examined. Secondly, the relationship between the number of Sci-Hub download requests in a geographic area and the number of residents in that area with advanced degrees was examined. This relationship was further explored by looking at how the presence of higher education institutions in that area changed this relationship. Examining the relationship between Sci-Hub usage, higher education, and advanced degree holders can inform whether the current academic literature system meets the needs of information seekers. While neither approach can definitively explain why people choose to use Sci-Hub, this study provides a better understanding of the phenomenon.

In order to frame the importance of Sci-Hub, this chapter examines the issues surrounding Sci-Hub and how it relates to the larger field of academic literature. This chapter begins with an examination of the history and purpose of academic literature and how it is and has been accessed. This chapter then discusses what Sci-Hub is, how it works, and how publishers have pushed back against it. Next, why studying access to academic literature is important, both functionally and theoretically, is explored. Lastly, the value and significance of this study within the field of higher education is explained.

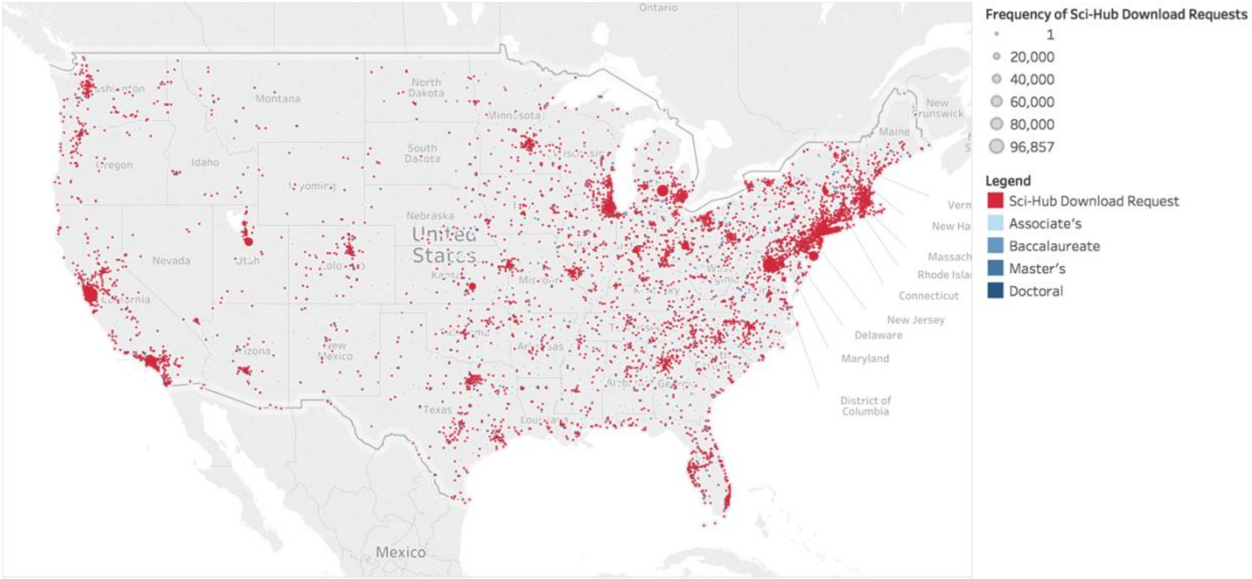

Figure 1. Sci-Hub Usage and Higher Education Institutions in the United States

This figure illustrates Sci-Hub download requests along with all degree-granting higher education institutions.

1.1 BACKGROUND ON ACADEMIC LITERATURE AND ACCESS

To understand the current academic publishing field and the reasons that Sci-Hub was developed, it is important to understand the history of academic literature and how access has developed and changed throughout that history.

1.1.1 A Brief History of Academic Literature

The scientific journal has existed for hundreds of years, beginning with the *Journal des sçavans* and *Philosophical Transactions of the Royal Society of London* in the mid-seventeenth century. Prior to the creation of journals, scientists would frequently keep their findings secret for fear they would be stolen; after gaining popularity, journals shifted this mindset and scientists began sharing their findings early and often to establish ownership and garner peer recognition. Over time, journals became more specialized and advances in printing allowed for greater dissemination. Most journals were the output of academic societies and the fees associated with subscriptions went to cover the costs of production and overhead for the society. While the distribution methods have changed over the centuries, the purpose of journals remains the same as when they began: a means of disseminating new theories and research to further the advancement of science and knowledge (Regazzi, 2015).

The growth in scholarly journal publishing mirrors the growth in higher education. In the United States, the G.I. Bill opened the path to a college education for many returning servicemen. As enrollments grew, new institutions were founded and existing institutions expanded; growing enrollments required more faculty members to meet the needs of colleges and universities. As the ranks of the professoriate grew, administrators looked to different metrics to gauge candidates and guide tenure and promotion decisions. The publish or perish model that emerged from these metrics created incentives for faculty to publish their research at greater and greater volumes (Greco, 2015) or risk being passed over for tenure and promotions.

1.1.2 Serials Crisis

Academics’ greater need for publishing articles, combined with the tradition of publishing these articles without direct financial remuneration, created an environment ripe for exploitation. New journals were launched that gave these new articles a forum, but many of these new publications were not the scholarly output of academic societies, but were instead created and maintained by commercial publishers who saw a profitable market. Academic libraries tried to keep apace of the new offerings, but rising journal prices accompanied by a rising volume of titles created what has become known as the serials crisis (Gennaro, 1977). The serials crisis, especially in the pre-digital publishing age, was borne out of multiple factors, including: a shift from academic or professional societies as publishers to commercial publishers, commercial publisher mergers decreasing competition, and the increase in titles and specializations of journals (Pascarelli, 1990).

The serials crisis was first identified in the 1970s when periodical price increases outpaced the consumer product index by over 200% (Gennaro, 1977). Concern about journal pricing and the increasing difficulties libraries had in maintaining their collections grew (Dougherty, 1989; Easton, 1999); as did journal prices and titles. By 1988, the number of journals had risen to over 40,000 (Broad, 1988) and serial expenditures rose by 227% between 1986 and 2002 (Association of Research Libraries, 2002). As digital publishing came to prominence, some scholarly societies like the American Chemical Society and the American Physical Society were established well enough that they were able to manage the format shift from print to electronic. However, many other disciplines were unable to make the change; society journals unable to make the shift to digital publishing were either acquired by commercial publishers or the societies made agreements with commercial publishers to provide a platform. This consolidation of content under a small number of commercial publishers coincided with sharp increases in publisher profits (Larivière, Haustein, & Mongeon, 2015).

In 1989, the Association of Research Libraries (ARL) laid out a series of options that could be taken to combat the serials crisis, including: education efforts to increase awareness in librarians, faculty, and administrators; identifying problem publishers and coordinating protests; resource-sharing; finding alternative, credible, non-commercial publishers; and working to reform academic promotion criteria to limit the pressure to excessively publish (Ivins, 1989). In 1998, ARL established the Scholarly Publishing and Academic Resources Coalition (SPARC) (Association of Research Libraries, n.d.); SPARC's mission is to "enable the open sharing of research outputs and educational materials in order to democratize access to knowledge, accelerate discovery, and increase the return on our investment in research and education" (SPARC, n.d.-b).

While frustrations with rising journal prices grew, the solutions were limited and some were fraught with their own perils. Libraries formed consortia to increase their bargaining power in attempts to negotiate lower journal prices with publishers (Wellen, 2004). The publishers' solution was to bundle journal titles and sell the bulk packages at a discount when compared to pricing per title. At first glance, journal bundling appears to offer libraries a respite from their financial duress, but a closer inspection uncovers less altruistic motives on the part of the publishers. The end result of these package deals is that high-demand journals are grouped together with titles that publishers might not otherwise be able to sell due to relatively low usage (Frazier, 2001; Wellcome Trust, 2003; Wellen, 2004). The bundles mean that while each title might come at a discount, some of the included titles might not otherwise be subscribed to. The true beneficiary of these deals are the publishers who have found a way to sell subscriptions to journals that would otherwise see little demand. While the greater access to information is desirable, the costs of these deals frequently forces libraries to make cuts elsewhere, resulting in, at best, a net neutral and likely a net negative.

1.1.3 Open Access

While librarians and other concerned participants in the academic literature field looked for ways to slow the crippling cost increases of academic journals while still providing students and faculty with access to the information they needed to perform their roles, it was the creation of the World Wide Web that provided the technical mechanism needed for cheap mass dissemination of information. As academics began to see the implications and possible usages of the Web, the open access movement began to emerge. Open access is free and unlimited access to scholarly works (Suber, 2015). There are several methods of achieving this goal which will be discussed further, but the ability for anyone to access scholarly output is the crux of the open access movement.

Many open access advocates argue that their aim is not new, it grows out of the scholarly tradition previously described where researchers and academics would present their findings in journals to further the advancement of science and knowledge (Regazzi, 2015); the difference is that the Internet has opened a new avenue for information sharing that can circumvent the costs of traditional print journals (“Read the Budapest open access initiative,” 2002; Willinsky, 2006). Although publishing open access journals stills requires some expenditures to cover technology and possibly editorial staff and typesetters, the costs are mostly limited to formatting and hosting the articles.

Groups such as SPARC, having already identified access and cost issues, quickly embraced the open access movement and have adopted it as one of their core missions (SPARC, n.d.-a). In Germany, a consortium of libraries, universities, and research institutes has formed with the goal of paying publishers an annual fee to access all of the publishers' content and have all articles with a German first author open to the world. While negotiations are ongoing and some publishers seem more hesitant than others, the consortium plans to hold firm and believes it will lower costs will increasing access (Vogel & Kupferschmidt, 2017).

As the open access movement has matured over the years, two main models have emerged: the gold model and the green model. The gold model involves publishing an article in an open access journal, or a hybrid journal that allows authors to choose whether the article is open or behind a paywall (Björk, 2016; Suber, 2015), while the green model involves placing articles in institution- or field-specific repositories. As previously stated, while open access journals can cost significantly less to publish than traditional journals, especially ones with print versions, they are not free to reproduce. A common method for cost recovery in open access journals is through an article processing charge (APC) which is paid by either the author or the author's institution, although some journal publishers have foregone the APC model for authors coming from institutions that pay an annual membership (Suber, 2015).

1.1.3.1 Open Access Journals (The Gold Model)

While open access has the ability to democratize access from the end user perspective, it is far from a panacea. Commercial publishers have found a business model that can maintain their profits if scholars choose to publish in their journals with high APCs. While the Managing Director for Scholarly Exchange, a not-for-profit publishing software provider, made claims as far back as 1998 of being able to publish an article with an APC between $50-$100 (Fisher, 2008), many of today's top commercial publishers charge a great deal more. Elsevier charges between $500-$5000 (Elsevier, 2016; "Open Access," n.d.) and Springer charges a flat rate of $3000 ("Springer Open Choice," n.d.). These high fees leave many researchers in less well-funded institutions or disciplines unable to afford placing their work in these types of open access journals.

Another barricade to more open sharing of research findings is the current publish or perish model of academic advancement. If institutions, departments, and administrators continue to use the same criteria to evaluate candidates for hire or promotion, then a large part of the problem that caused the serials crisis will continue. If a few journals in a discipline are considered to be the gold standard by tenure committees and administrators and those journals are owned by for-profit, commercial publishers, then the journals can continue to charge exorbitant prices whether those come via subscription or APC. In essence, these departments and schools are a major contributor to the vicious cycle that keeps journal prices skyrocketing at rates that far outpace inflation.

1.1.3.2 Repositories (The Green Model)

The green model of open access revolves around repositories, generally either discipline- or institution-specific repositories. Repositories can contain either post-print or pre-print versions of articles. Post-print articles have been peer reviewed and accepted for publication; pre-print articles are generally the version that is initially submitted for publication, but has not yet been peer reviewed. Post-print repositories are either institutionally run, like Harvard's DASH and the University of Pittsburgh's D-Scholarship@Pitt (D-Scholarship@Pitt, n.d.; Office for Scholarly Communication, n.d.), or subject-specific, like PubMed Central and the NSF Public Access Repository (National Library of Medicine, n.d.; National Science Foundation, n.d.-a). Some repositories are optional; however, funders such as the National Institutes of Health and the National Science Foundation require articles published stemming from their research grants to be uploaded to the repository within a certain timeframe from the date of publication (National Science Foundation, n.d.-b; NIH, n.d.-b).

Pre-print repositories began in 1991 with the creation of arXiv, a repository based at the Cornell University Library that focuses on physics, mathematics, and similar fields. Articles hosted in arXiv are moderated to “verify that they are topical and refereeable scientific contributions that follow accepted standards of scholarly communication” (arXiv, n.d.), but they are not peer reviewed. While arXiv has been around for decades, there has been a relatively recent growth in the number of pre-print repositories and their popularity. Services such as F1000Research, bioRxiv, SocArXiv, and PeerJ Preprints have come along with their own subject-specific pre-print repositories (Cold Spring Harbor Laboratory, n.d.; F1000Research, n.d.; PeerJ, n.d.; SocArXiv, n.d.). One of the top criticisms of these services is that pre-print repositories, as previously mentioned, do not provide peer-review for the articles prior to making them publicly available. However, as arXiv advocates point out, traditional peer-review does not certify research results or protect against fraud and the overhead costs and delay in publication make timely access to findings difficult (Gunnarsdóttir, 2005).

1.1.4 Interlibrary Loan

Interlibrary loan (ILL), also known as document delivery, is the process that libraries use to request materials that they do not possess from other libraries for patron use (ALA, n.d.). Libraries can be both borrowers and lenders in this process and often form consortia and partnerships (Shrauger & Scharf, n.d.). Borrowing libraries pay a fee to the lending library to offset the costs of sending the material. This method works to help libraries with large collections offset some of their expenditures and for libraries with smaller collections to meet their users' needs without exceeding their collections budgets.

While this process worked well for physical journal articles that would be otherwise inaccessible, and still works well for physical materials, it creates cumbersome time and process delays for users seeking electronic journal articles. As an example, article requests at the University of Pittsburgh can take four days to be completed (Colbert, n.d.) and students at Grove City College need a signed form from a professor before an article request will be processed (Cavanaugh, n.d.). The ILL process, while technically capable of meeting users' information needs, creates delays in access to information that exist solely because information is paywalled.

1.1.5 Article Sharing

A more informal method of article dissemination is through articles being shared by the authors. People interested in reading an article they do not have access to have long been able to write to the authors and ask for a copy, even when that method involved sending a letter and hoping to receive a copy through the mail. The rise of social media, including academia-focused social media sites like ResearchGate ("ResearchGate," n.d.) and Academia.edu ("Academia.edu," n.d.), have made it easier for potential readers to connect with authors. While these sites have encountered legal issues when authors post their articles there (Chawla, 2017), the sites can still be used to send messages to authors to ask for articles to be sent via email.

Another recent development in access is article sharing through publishers. SharedIt, available through publishing giant Springer, allows authors to send a link to an article to collaborators and other interested parties, including by posting on social media. While the SharedIt program provide an alternative method to the traditional open access methods for articles that would otherwise be behind a paywall , it does not allow for these articles to be printed or downloaded ("Principles and guidelines," 2016). Additionally, this requires the authors to post the link somewhere that potential readers can find it or respond to email inquiries asking for access. While better than providing no access method beyond paying a fee, there are still limitations to access and use.

1.1.6 Guerilla Open Access

Guerilla Open Access is a term coined by activist Aaron Swartz (2008) to describe the process through which individuals with access to academic literature can, and should, share otherwise paywalled information with the masses. For Swartz, the act of sharing these materials may be illegal, but justifiable. "It's called stealing or piracy, as if sharing a wealth of knowledge were the moral equivalent of plundering a ship and murdering its crew. But sharing isn't immoral — it's a moral imperative." (Swartz, 2008). For the purposes of this study, Guerilla Open Access (GOA) will be defined as a method of providing access to paywalled information that circumvents intentional access limitations, regardless of legality.

While Sci-Hub may be the most well-known and largest example of GOA, it is certainly not the only example. Swartz himself downloaded 4.8 million articles from the journal database JSTOR using the Massachusetts Institute of Technology's network. The results of Swartz's downloads and the discovery of his act led to Swartz being charged by the federal government under the Computer Fraud and Abuse Act. Swartz hanged himself before going to trial (Bombardieri, 2014). Thankfully, most acts of GOA do not end in tragedy.

The Scholar subreddit ("Scholar," n.d.) is a forum that allows users to request a document by providing the DOI, PMID, or ISBN. Other users, known as fulfillers, will then retrieve a copy of the file and place a link to it in the comments (RoyalKoala23, n.d.). The #ICanHazPDF hashtag on Twitter follows a similar small-scale method of GOA where an individual makes a request and another individual procures the article and shares it. In this method, requestors tweet an article title, DOI, or some other unique identifier with the #ICanHazPDF hashtag and their email. A second user will download a copy of the article and email it to the user; upon receipt, requestors are encouraged to delete the tweet to conceal the transaction (C. Gardner & Gardner, 2015). A study of these crowdsourced types of GOA found that users' primary motivation is utilitarian, these methods are faster than interlibrary loan (C. C. Gardner & Gardner, 2016).

1.2 SCI-HUB

While smaller one-to-one examples of GOA provide some relief to information seekers stymied by paywalls, they are not scalable to a degree that would meet the information needs of everyone. Sci-Hub declares itself to be "the first website in the world to provide mass & public access to research papers" ("Sci-Hub," n.d.). To explain why Sci-Hub is worthy of study, it is important to understand what Sci-Hub is, how it works, and how publishers have responded to it.

1.2.1 What is Sci-Hub?

Sci-Hub is a web service with an associated repository. The aim of Sci-Hub is to provide free access to academic literature through a single, simple search interface. The site hosts and serves out articles regardless of their copyright status. To better understand why Sci-Hub was launched, its creator, Alexandra Elbakyan (2015), describes her rationale for developing Sci-Hub:

> I would like to clarify the reasons behind sci-hub.org website. When I was a student in Kazakhstan university, I did not have access to any research papers. These papers I needed for my research project. Payment of 32 dollars is just insane when you need to skim or read tens or hundreds of these papers to do research. I obtained these papers by pirating them. Later I found there are lots and lots of researchers (not even students, but university researchers) just like me, especially in developing countries. They created online communities (forums) to solve this problem. I was an active participant in one of such communities in Russia. Here anyone who needs research paper, but cannot pay for it, could place a request and other members who can obtain the paper will send it for free by email. I could obtain any paper by pirating it, so I solved many requests and people always were very grateful for my help. After that, I created sci-hub.org website that simply makes this process automatic and the website immediately became popular. (para. 3)

Essentially, Elbakyan experienced the same access limitations outlined previously and took actions designed to circumvent these limitations.

When users search for an article on Sci-Hub, a search of the repository is performed. If the article is stored in the repository, the user is taken to the hosted copy of the article. However, if the article is not in the repository, Sci-Hub uses the credentials of someone at an institution with access to obtain a copy of the article; this copy is both presented to the requestor and stored in the repository for future use (Cabanac, 2016). As these copies are taken directly from the publishers' platforms, an article obtained via Sci-Hub is no different for the reader than a copy attained legally, unlike articles in pre-print repositories which may differ slightly from the final version. Additionally, this process helps ensure that the most popular articles are in the repository, since it is the demand of users that feeds the repository.

Since its launch, Sci-Hub has generated a lot of web traffic. An analysis of the server logs during the six-month period from September 2015 through February 2016 shows that 28 million documents were served out to users all over the world (Bohannon, 2016b). *The Chronicle of Higher Education* reached out to the authors of Sci-Hub's most downloaded articles for their reactions and found they were generally pleased with the increased exposure, citations, and dissemination, although one was worried about the long-term effects on society journals (Ruff, 2016). Elbakyan (2015) adds that Sci-Hub has "never received any complaints from authors or researchers, only Elsevier is complaining about free distribution of knowledge" (p.2). From an access perspective, Sci-Hub has created a method of sharing information beneficial to both authors and readers.

Sci-Hub, while important for meeting the needs of information seekers, is not a panacea. At its core, Sci-Hub is a repository of materials generated using other means and does not provide for peer-review and some of the other important work that goes into creating academic literature. Sci-Hub is also primarily run by one person and the functions of the site are susceptible to the whims of Elbakyan. In September 2017, Sci-Hub announced that it would be blocking the site to Russian users (Standish, 2017). In a letter posted on the Sci-Hub homepage at the time, Elbakyan states the decision was based on what she deemed offensive behavior by Russian scientists toward her, including naming a parasitic insect after her and alleging that she is insane. She did, however, offer suggestions on other methods of accessing information to those in need (McLaughlin, 2017). While her stance and this maneuver may be understandable, it also highlights the need to look at how information needs are met, or not, in the long term.

1.2.2 How Publishers Fight Sci-Hub

Unsurprisingly, the academic publishing companies who hold the copyright of the articles hosted in Sci-Hub have pushed back. Elsevier and the American Chemical Society (Association of American Publishers, 2017; Schiermeier, 2017) have filed suit against Sci-Hub in the American courts and won. As a result of Elsevier's suit, Sci-Hub had to undergo a domain name change; however, as Sci-Hub is based in Russia, these lawsuits have been unable to shut the site down (Bohannon, 2016b). Not only have these lawsuits been ineffective in stopping Sci-Hub, there is evidence that each legal challenge resulted in an increase in Google searches for the Sci-Hub website (McKenzie, 2017).

In addition to the lawsuits, publishers have worked in concert with colleges and universities to block the methods used by Sci-Hub to garner access to articles. One such method to thwart Sci-Hub from accessing an institution's resources is to require two-factor authentication to access copyrighted content from off-campus. Two-factor authentication, as the name implies, requires two separate sets of credentials, to authenticate access. Common forms of authentication include username and password, ID and PIN, and software tokens (Elsevier, n.d.); frequently two-factor authentication manifests itself by having credentials sent to smart phones for confirmation ("Multifactor Authentication at Pitt," n.d.). Two-factor authentication, while creating extra work for users, is intended to prevent credential sharing. Where access rules are circumvented, publishers will also cut off access to resources for an institution if there is evidence of massive downloading. In my professional role as a systems librarian, I have had to assist in tracking down information on accounts associated with excessive downloads. Institutional access is restored once the offending account has been disabled.

Publishers are so concerned with the threat posed by Sci-Hub that they have even taken to attacking academics who present information about it. In 2016, Gabriel Gardner, a librarian at California State University, Long Beach, discussed Sci-Hub as part of a conference panel on resource sharing and the future of interlibrary loan. As a result of this discussion, the president of the Association of American Publishers wrote a letter to Gardner's dean admonishing Gardner, framing his comments as supporting Sci-Hub as opposed to explaining it. While the dean sided with Gardner (Jaschik, 2016; Masnick, 2016; Peet, 2016), the implied intent of this intimidation is to keep academics from even discussing Sci-Hub.

In an interview about the findings of a study on Sci-Hub examined in more detail in the next chapter (McKenzie, 2017), Himmelstein adds that new technologies could allow papers to be hosted in a manner that would not be centrally located, which would make it nearly impossible to shut down a service like Sci-Hub. He added,

> I think the larger picture of this study is that this is the beginning of the end for subscription scholarly publishing. I think it is at this point inevitable that the subscription model is going to fail and more open models will be necessitated. One motivation for doing the study is that I want to bring that eventuality into reality more quickly. (McKenzie, 2017).

As publishers continue to impose roadblocks on Sci-Hub usage, the rapidly changing technological environment will continue to make that more difficult.

1.3 IMPORTANCE OF STUDYING ACCESS TO ACADEMIC LITERATURE

Research findings and theories, the primary content of academic literature, on their own do nothing; it is through sharing that they have value. Therefore, a study of academic literature must also examine access to the literature. The function of academic literature is to share with others, it is through that process that new theories and lines of inquiry are developed, scientific progress is made, and knowledge is disseminated. Academic literature also serves as the product created by scholars in their role as workers in higher education and research institutions. It is this role of academic literature that necessitates examination through a critical theory lens.

1.3.1 The Functional Importance of Access to Academic Literature

College and university faculty are the most frequently mentioned group of people affected by access limitations. The primary reason for this is the faculty's role in research, although keeping abreast of developments in their respective fields for teaching purposes is also vital. Access to academic research provides an example of a virtuous cycle. The more open and available research findings are to people, the greater the opportunities to build off this research and make further scientific gains. These new findings are then shared with the community and the cycle continues. Berry (2001), a chemistry professor at the University of Chicago, frames it as, "These are goods whose value does not diminish with use. In fact, because science functions in a cumulative way, building on previous knowledge, the more the results are used, the greater is their value." (p. 38). Not only are journal articles non-rivalrous goods in that reading one does not diminish its value for the next reader, the knowledge garnered can be utilized to further build upon it. By creating easier pathways for researchers to share their work and absorb the work of others, there is a clear benefit to all parties involved.

For individual faculty, access to academic literature can also be personally beneficial. From the late 1970s through 2001, Tenopir and King (2001) surveyed nearly 15,000 scientists to track their communication and reading habits. One of their most interesting findings is the relationship between journal article reading and professional success; award winners read 53% more than non-award winners and scientists considered high-achievers by their peers read 59% more than their colleagues when holding other variables constant. These findings help to underscore the value of access to academic literature. Professional advancement and notoriety is another reason why access matters to faculty. One important factor in faculty tenure and promotion decisions is how often publications have been cited. Fewer access restrictions on an article means a wider audience for the work which can increase the likelihood of an article being cited (Suber, 2015).

For faculty and researchers in the developing world, access to information is crucial and frequently hard to come by. Much like their counterparts in wealthier countries, faculty and researchers in developing nations need access to research to improve the health, food supply, environment, and policies of their countries. The United Nations recognized this need and over the past decades developed a series of public-private partnerships known collectively as Research4Life. Research4Life includes HINARI, AGORA, OARE, and ARDI representing health, agriculture, environmental studies, and science and technology, respectively. These programs, in conjunction with journal publishers, provide free or low cost access to developing countries (“Research4Life home,” n.d.). Greater access for these scholars can lead to real, tangible, and immediate benefits to their societies.

The virtuous cycle of sharing research findings and theory with the other members of the academic community, combined with the personal benefits that more universal access can provide, help to demonstrate why access to academic literature is important for faculty and researchers. However, faculty are not the only group of people who benefit from greater access to academic literature. Graduate students, in their dual roles as students and budding researchers, may face the greatest challenges in terms of access. In their role as students, they need their faculty members to have access to all recent developments in the field so that information can be shared and discussed. Secondly, as researchers, graduate students require access to academic literature in similar ways that faculty, as researchers, need it. While most undergraduates do not participate in research in the same way as their peers in graduate programs, they can still be stymied when trying to complete their coursework.

Additionally, the public benefits from greater access to academic literature. With the recent glut of misinformation and propaganda in the public sphere (Timberg, 2016; Wingfield, Isaac, & Benner, 2016), there is a need for the public to have access to high quality, accurate information. While academic literature is not written with the layman as the intended audience, barriers to access to this literature creates a need for the results to be filtered through a third party that is subject to its own agendas. These third parties can change the message of the research for myriad reasons ranging from the relatively benign of trying to make them more sensational to get more readers (Nolan, 2012; Whiteside & Hardin, 2011) to purposeful mischaracterization for the sake of changing public opinion, or at least sowing the seeds of doubt, on issues like climate change (Harvey, 2016; Lewandowsky, Ballard, Oberauer, & Benestad, 2016). While not every individual will have the desire to read academic literature, there will be some who do and who can help their fellow citizens combat misinformation. The value of academic literature for research and teaching purposes is clear, but, in this form, it can also function as the third piece of the academic triad: service.

It is important to understand the benefits of greater access to academic literature, but understanding the deleterious effects of access limitations is also essential to understand the importance of studying the subject. Access limitations can be placed into two categories: inability to access and access hindrances. The inability to access academic literature is, as the name suggests, having no legal method to retrieve needed information. While this form of access limitation is the most difficult to overcome, the second form is also pernicious. Access hindrances mean that while there are methods to acquire the desired literature, the path is often cumbersome, time consuming, and requires knowledge of the procedures to request access.

For those with no ability to access academic literature, it is likely that they are not associated with a higher education institution, hospital, or research institution or they are in a developing nation that does not have interlibrary loan agreements. For members of the public in the United States, this type of access limitation could include teachers looking to stay current with trends in both the field of teaching and in their subject area, health professionals at clinics without an affiliation with a large hospital system, or a member of the general public researching a topic meaningful to them. This could mean lower quality of teaching, healthcare that does not meet the most recent protocols, or the inability to be informed about a topic relating to themselves or their family.

Internationally, while the aforementioned Research4Life program ("Research4Life home," n.d.) provides access to journals in certain fields for developing nations, these programs do not include education or the humanities. While the subject-areas covered may be more pressing for survival, they do not offer the full breadth of research available in wealthier nations, exacerbating existing inequalities. Additionally, these programs are contingent on the continued participation of the publishers and can be revoked at any time (Z. Kmietowicz, 2011), making access tenuous.

Access hindrances, while less severe than their counterparts, are also problematic for a number of reasons. Without the institutional knowledge of how to circumvent these hindrances, the hindrance becomes a de facto inability to access. An undergraduate student preparing to apply to a PhD history program describes their experience with access limitations:

> ‘The library at [CSU attended] does not have a lot of resources for students who want to conduct research work in history. My research paper [the one I wrote about in my statement of purpose and submitted as a writing sample] was not as good as it could have been because of the resources available at (CSU attended) were so limited.’ (DeAngelo, 2010, pp. 27–28)

While it is likely that this student could have retrieved the desired literature through the university’s interlibrary loan program, that process could have added costly time delays to the application process and if the student was unaware of the program, this hindrance becomes a roadblock.

Additionally, these access hindrances work against human behavior. Zipf’s Principle of Least Effort (1949), described in greater detail in chapter two, posits that people naturally gravitate towards their desired outcome along a path that provides the most immediate solution that does not simultaneously create long term problems greater than the problem at hand. This theory has been applied to information seeking in higher education and a study suggests that information seekers in higher education are looking for the best answer in the least amount of time and with the least effort expended (Connaway, Dickey, & Radford, 2011). While information seekers may technically have access to an article, the delays involved with interlibrary loan may lead to the use of a different article, even if it does not meet the information need as well. This could also explain why some interlibrary loan requests are never retrieved even after they have been filled (Shrauger & Scharf, n.d.).

1.3.2 The Theoretical Importance of Access to Academic Literature

To this point, the functional purposes of academic literature and how access affects academia as a whole and the individuals who comprise it have been discussed. However, there is also a need to step back and look at academic literature through another lens. Critical theory embraces the critical method inherent in Marxism and focuses it on the political and cultural structures of society. Instead of focusing on a rigid system of thought, critical theory focuses on liberation from current forces of oppression (Bronner, 2011). A key component of both traditional Marxism and critical theory is historical materialism. Historical materialism places facts within their historical context as products of social action. The economy, the state, and culture are the primary components of the totality, the all-encompassing social relations that shape the world. Within this framework, moments and movements are recognized as being both influenced by and influencing the totality (Sherman, 2016). Critical theory provides another framework for understanding why studying the production and dissemination of academic literature is important.

Alienation and reification are two common, intertwined themes in critical theory. Alienation, in critical theory, is the separation of the person from the product they create. With its basis in Marxism, alienation often reflects the concerns of the assembly line worker whose labor creates one small part of a larger item and, thus, divorces the worker from the finished product and distances workers from each other and their final products. The resultant alienation from the finished products of labor results in a lack of fulfillment and misery (Fromm & Marx, 1966). Reification is the idea of reducing a person to a part or role; the worker is simply one more piece of equipment in the assembly line. For example, the capitalist would look at a worker as a welder, not a person who welds as part of the larger project (Bronner, 2011; M. Peters, Lankshear, & Olssen, 2003). Horkhemier (1982) posits that the elevated social position given to scientists leads them to believe that they do not fall within this structure, but without a critical theory frame the work of scholars and scientists reinforces the status quo and perpetuates the process of recreating the hegemony. In other words, as Fromm states in Marx's Concept of Man (1966), "Intellectual activity is of course, for Marx, always work, like manual or artistic activity." (p. 47). While faculty and researchers may not work on an assembly line, they still work in the highly segmented realm of higher education and are still subject to the forces of alienation and reification.

Historical materialism is the totality of social relations, the combination and coordination of political, economic, and cultural forces. Critical theory, relying on historical materialism, requires placing issues in the context of their time and place. In an era heralded as the Information Age (Birkinshaw, 2014) or within a knowledge economy (Powell & Snellman, 2004), access to information becomes a key struggle. Researchers, faculty, and students become both producers of information and users of it. Access to information becomes a privilege for those who can afford it or can align themselves with an organization that can afford it, thus exacerbating existing gaps.

Reification, as outlined above, is the act of turning people into things. Recently, the singular thing that is frequently used to describe people is as a consumer. By framing people as consumers, the neoliberal hegemony shifts all human interaction into commercial transactions. If an academic article exists, the author is the producer and the reader is the consumer or perhaps more accurately, the author is the laborer who creates the material, the publisher is the capitalist who sells the article, and the reader is the consumer.

Furthermore, in this capitalistic formation of knowledge creation and dissemination, academics are put into competition with one another. Their work becomes not a piece of a greater whole dedicated to furthering human knowledge, but a steppingstone to greater individual glory. The number of articles published, the number of citations these articles generate, and the perceived value of the journals where these articles are published lead to tenure and career advancement. With these goals in mind, colloquially known as publish or perish, researchers become divorced or alienated from their work as a summation of their findings and focus on what these findings can do for them. The need for greater numbers of published articles creates a vicious cycle where journal publishers create new journals to accommodate this perceived need. Publishers then sell these journals back to the institutions that fund the researchers, generating ever-increasing profits. However, critical theory challenges the assumption that academic literature must follow this market orientation.

Pyati (2006) applies the works of critical theorist Herbert Marcuse to the field of information studies through Marcuse's focus on technological rationality. Technological rationality is the idea that advanced societies make scientific and technical progress into instruments of domination. For Pyati, in this techno-capitalist framework, "Information, in its modern sense, became dissociated from affective, contextual, and cultural processes, thus making it much easier to be commodified, reified, and abstracted." (p. 85). By removing academic literature from the process of scientific discovery and theorizing and divorcing it from the authors creating it, the literature becomes an abstract object that can be repackaged and sold in whatever manner is most profitable.

Critical theory is not only a framework for critiquing societal ills, but also necessitates praxis for actively challenging the status quo. A decade ago, Pyati (2007a) examined open access through a critical theory lens and found a number of methods used to combat the serials crisis. He noted SPARC, described earlier, as a coalition dedicated to reducing the economic hardships caused by rapidly increasing journal prices. Libraries and librarians are a major focus of Pyati's article; specifically, their role in promoting, and even creating, open access journals along with their work with repositories. While these methods are commendable for having helped slow cost increases, the underlying problems persist.

From a critical theory perspective, Sci-Hub can be viewed as tool for subverting the for-profit publishing model. Sci-Hub shows a direct action taken to make access to information more universal; less concerned with the legality imposed by the powerful and more concerned with the morality of universal access, Elbakyan has taken a stand against the status quo. This stand is not without risk as Elbakyan "is at risk of financial ruin, extradition, and imprisonment because of a lawsuit launched by Elsevier" (Bohannon, 2016b). While critics of some of the methods employed by Sci-Hub will cite the extralegal or possibly illegal activities, critical theory is concerned less with what is legal and more with what is right and just.

Understanding the importance of access to academic literature falls into two categories. For the functional purpose of academic literature, the benefits of greater access and the deleterious effects of access limitations demonstrate how valuable access is. Knowing what access looks like is vital to understanding what needs to be done to provide more universal access. For the producers of academic literature, it is important to understand how access limitations have been used to commodify knowledge and divorce it from its creators. Through this greater understanding, steps can be taken to make researchers part of a community that builds upon each other's work for the betterment of all instead of in competition with each other for personal gain.

1.4 SIGNIFICANCE AND PURPOSE OF THE STUDY

At the heart of the debate around the use of Sci-Hub, especially in the United States, is the question of whether academic literature is a public good, designed to share information and knowledge, or a commodity, a product to be sold like any other consumer product. While the open access movement has presented the academic publishing industry with an alternative method for sharing journal articles, the industry has still found a way to profiteer from the process. Sci-Hub, however, represents the first large-scale disruptive force to challenge publishers' hegemonic control. The relative newness of Sci-Hub and its legally questionable methods have contributed to limited studies regarding its use. This study sought to better understand how Sci-Hub usage in the United States is related to higher education and the educational attainment of population centers.

In the commodity view of academic literature, creating a market and profiting from the publication of scientific research is not only reasonable, it is right and just. Proponents of the for-profit publishing model, regardless of the method of revenue generation (traditional reader-pays model or APCs), argue that the industry has not restricted growth in knowledge creation and sharing, but has actually increased the speed of development (Jongejan, 2003). The purchasers of this product, primarily academic libraries, make choices to subscribe to the academic journals that best fit the perceived needs of their patrons. Publishers further assert that if a library's user needs access to an article from a journal that the library does not subscribe to, there are legal methods in place for acquiring that article (McNutt, 2016) and that use of Sci-Hub in countries like the United States is based on convenience, not necessity (Bohannon, 2016b).

The clustering of Sci-Hub download requests in the United States near the locations of research institutions (see Figure 2) suggests support for the convenience over necessity argument. However, there has been no previous analysis comparing the journal collections of academic libraries with the Sci-Hub download requests in the United States from the areas surrounding these institutions. One purpose of this study was to look at the relationship between academic library journal expenditures and the number of Sci-Hub download requests near said library. Additionally, a post-hoc analysis was performed that took a sample of Sci-Hub download requests, ascertained the journal that published the article, and searched the nearby academic library's catalog to determine if the library subscribes to that journal.

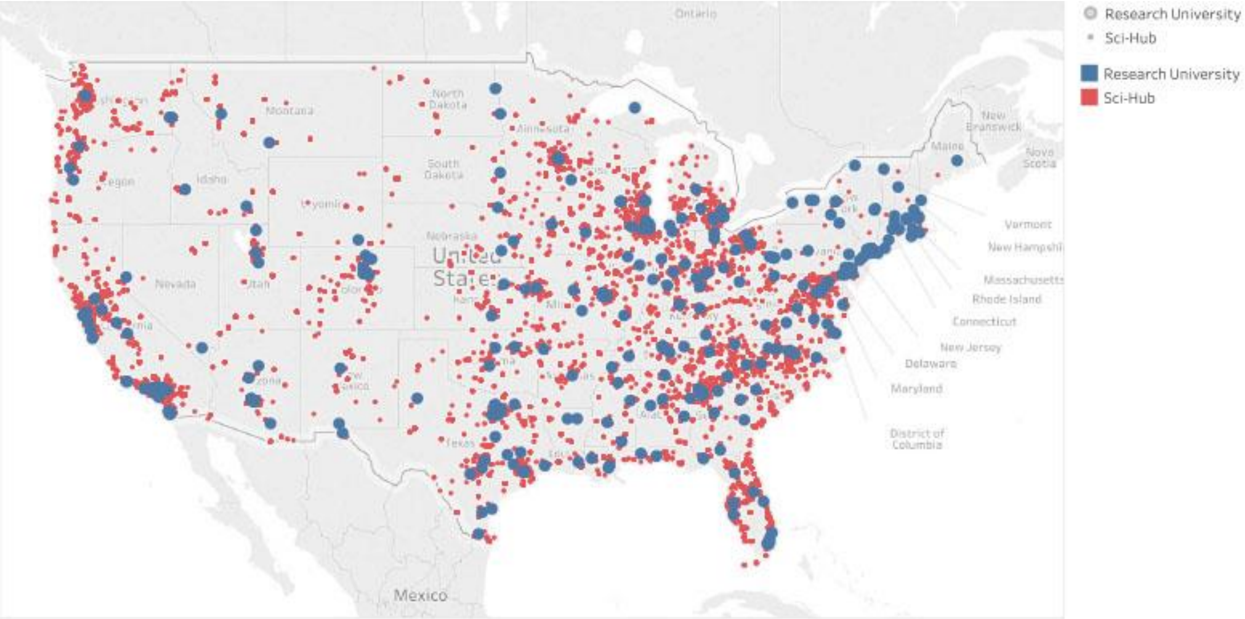

Figure 2. Research Universities and Sci-Hub Download Requests

This figure shows the location of the 335 research universities and all Sci-Hub download requests in the United States.

A second set of research questions examined the relationship between Sci-Hub download requests and the population of the geographic area where the request was generated (see Figure 3). As academic literature is primarily written for members of higher education institutions and professionals, an analysis was performed to better understand the relationship between the number of Sci-Hub download requests in a geographic area and the percentage of the population with an advanced degree, specifically with a Master's degree or higher. Further analysis was done to account for the number of higher education faculty members, the number of graduate students, and the number and type of higher education institutions in the region.

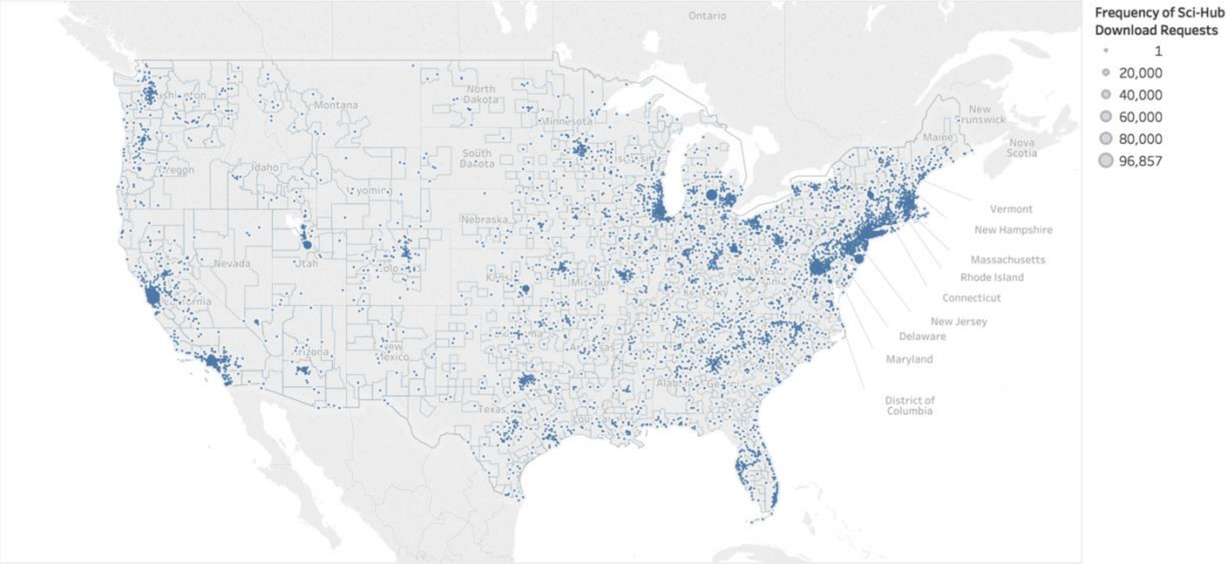

Figure 3. Sci-Hub Download Requests by CBSA

This figure shows Sci-Hub download requests overlaid on a U.S. map divided by CBSAs.

1.4.1 Delimitations

While a definitive answer to who is using Sci-Hub and why would be helpful for understanding Sci-Hub's role in information dissemination in the United States, this study cannot provide that. The nature and content of the datasets forces some assumptions that, while likely as a whole, probably have some exceptions. For instance, a Sci-Hub download request from near a higher education institution cannot indicate if the user is, in fact, associated with that institution. Also, the rise in online higher education programs means that the geographic location of a person is less indicative of where they attend college.

As outlined previously in this chapter, different academic disciplines have their own protocols and practices regarding information sharing. As an example, the fields of physics and mathematics have been using arXiv as a repository for decades to share findings (arXiv, n.d.); however, not every discipline is as open. Likely as a result of disciplinary differences, some fields tend to be overrepresented in Sci-Hub (Greshake, 2017a; Himmelstein, Romero, McLaughlin, Tzovaras, & Greene, 2017); while these disciplinary differences are addressed in the next chapter, this study did not explore differences in Sci-Hub use between disciplines.

Lastly, Sci-Hub, and Guerilla Open Access more generally, may represent a disruptive force beyond access limitations, but that is beyond the scope of this study. For instance, the commodification of information helps frame why Sci-Hub exists, but this study did not address the commodification process or how Sci-Hub may change that process.

1.4.2 Summary

Studying the relationship between Sci-Hub download requests, the various components of higher education outlined above, and advanced degree holders cannot definitively demonstrate whether or not use of Sci-Hub is for convenience or necessity, but it can help to better understand the phenomenon. This study was not designed to extoll the virtues of Sci-Hub, nor to condemn its circumvention of copyright laws. The rationale for studying Sci-Hub use in the United States is to better understand how members of the higher education community and the general population deal with access limitations when seeking information to meet their needs.

2.0 CONCEPTUAL FOUNDATION AND LITERATURE

This study, explored the relationship between academic literature and the people who use it. In this chapter, the Principle of Least Effort is used as a way to understand information seeking behavior and how that applies to academic literature. How Sci-Hub has been previously studied, what gaps in the literature exist, and how this study helps to fill those gaps are also explored.

2.1 CONCEPTUAL FOUNDATION

Sci-Hub, at its core, is a tool. It is used as a method of obtaining access to desired information. It is highly unlikely that someone would go to Sci-Hub without an information need; as such, it is important to understand how people seek information. Much of the literature on information seeking behavior centers on how people search for information, based on the nature of their information needs, and how they decide which information sources present the best answers to their questions (Case & Given, 2016; Choo, Detlor, & Turnbull, 2000). However, the problem that Sci-Hub seeks to solve, or at least circumvent, is not a question of finding information, but accessing it. From this position, the method a person uses to find an article that will meet their information needs is immaterial; the method they use to acquire it is paramount.

2.1.1 Zipf's Principle of Least Effort

Zipf's (1949) Principle of Least Effort (PLE) posits, amongst other things, that people will naturally choose the path of least resistance or effort.

> In simple terms, the Principle of Least Effort means, for example, that a person in solving his immediate problems will view these against the background of his probable future problems, *as estimated by himself*. Moreover he will strive to solve his problems in such a way as to minimize the *total work* that he must expend in solving *both* his immediate problems *and* his probable future problems. (Zipf, 1949, p. 1).

In short, the PLE suggests that people naturally gravitate towards their desired outcome along a path that provides the most immediate solution that does not simultaneously create long term problems greater than the problem at hand.

PLE is frequently applied to the field of information seeking. A 2015 study of library and information science articles published between 1949 and 2013 that cited Zipf's *Human Behavior and the Principle of Least Effort* (1949) showed that nearly a quarter of these article reference PLE; nearly 65% of the articles reference Zipf's Law, which refers to the frequency of word usage and is typically applied to bibliometrics. The trend lines for referencing PLE decreased throughout the 1970s and 80s before hitting a low point in the early 1990s; the trend started to increase again in the mid 1990s (Chang, 2016). The positive change in the trend line occurred at approximately the same time as the rise of the World Wide Web. Prior to the advent and popularization of the Internet and the World Wide Web, information seeking was a process of finding the best possible answer within the limits of what information sources were available. Frequently, this meant using a local library and utilizing the organizational and information finding practices of the library. Now with a glut of information at most people's fingertips, the process is more a matter of finding the best answer, or at minimum an answer that will meet their needs, in the least amount of time.

Connaway, Dickey, and Radford (2011) studied faculty, undergraduates, and graduate students at American colleges and universities and found convenience to be central to information seeking behavior and especially so for the millennial participants. For their study, convenience was broken down as finding sources that were good enough to satisfice their information needs, ease of access to the information source, and the amount of time spent finding the information. Based on their results, it can be inferred that information seekers in higher education are looking for the best answer in the least amount of time and with the least effort expended. In terms of Sci-Hub usage, the ease of access to the information source is the key component of convenience as defined in the study.

Schwieder (2016) believes that PLE is so central to information seeking behavior for academic library users that he developed a toolkit to assist users with low-effort information seeking strategies. Specifically, Schwieder advocates for a heuristic information seeking approach that uses a dual process approach that combines PLE with other simple best practices to account for the quality of information along with the ease of access. One of Schwieder's recommendations is the use of Google Scholar based on the simple search box and consistency. While Google Scholar can be configured to implement links to library resources ("Google Scholar Support for Libraries," n.d.), browsers can also be configured to go directly to Sci-Hub (Marcos, 2017; "The Installation of Sci-hub Plugin," n.d.). Each method requires approximately the same effort level to install, but the Sci-Hub plugins provide access to more resources.

The application of PLE to Sci-Hub provides additional insight into understanding why the system is utilized. As described in Chapter 1, the rise in popularity of Sci-Hub has seen colleges and universities work with publishers to combat unauthorized access. By understanding Zipf's Principle of Least Effort, we can envision how some of these methods, especially two-factor authentication, can actually increase Sci-Hub usage for materials available through institutional licenses as the prescribed methods become more laborious.

For example, a scholar reviewing literature off-campus may be unable to access their smart phone. If the scholar's library subscribes to the needed journals, the choice for this scholar is now to either wait until they are on campus, wait until they have access to their phone to complete two-factor authentication, or use an extralegal method for accessing the literature. Based on PLE, the clear choice is to continue working, even if that means accessing the needed articles through Sci-Hub. If the scholar's library does not subscribe to the needed journals, the methods for accessing these materials, such as interlibrary loan, can take days to deliver the desired articles and part of Connaway et al.'s (2011) definition of convenience relies on timeliness. In this second scenario, the choice to use Sci-Hub or some other method of GOA becomes even more tantalizing. In an interview, Sci-Hub researcher Himmelstein responds to a question about what publishers could do to stop new articles from being added to Sci-Hub: "There are things they could do but they can really backfire terribly. The issue is the more protective the publishers are, the more difficult they make legitimate access, and that could drive people to use Sci-Hub." (McKenzie, 2017).

If the purpose of academic literature is to share scientific discoveries and theories and users seek information based on a combination of best quality with least effort, then Sci-Hub becomes the choice that delivers the most comprehensive results with minimal effort. Again, this perspective on why people may use Sci-Hub in the United States lends credence to the idea that it is more a matter of convenience (Bohannon, 2016b); for Sci-Hub users affiliated with academic institutions, there may be methods of acquiring needed materials (McNutt, 2016), even though they run counter to how users seek information (Connaway et al., 2011; Zipf, 1949). However, with the pressures to conduct research and publish (Liebowitz, 2015), circumventing inconveniences may also be viewed as a necessity.

For the purposes of this study, convenience is defined as using Sci-Hub to download an article that would be available for immediate download through an affiliation with an institution that licenses that resource. This would exclude any article that would be available via interlibrary loan as this process violates the ease of access to information and timeliness portion of Connaway et al.'s definition of convenience. Conversely, in this study, necessity is defined as using Sci-Hub to download an article that would be otherwise unavailable for immediate access without paying a direct fee to the publisher.

2.2 REVIEW OF RELEVANT LITERATURE

The literature on Sci-Hub falls into two categories: how Sci-Hub is viewed and how Sci-Hub is used. The former uses arguments based on legality, morality, and functionality. The latter examines the data provided by Sci-Hub to better understand how Sci-Hub is used and attempts to infer motivations based on the results of data analysis.

2.2.1 Views on Sci-Hub

Sci-Hub represents a serious and direct disruption to the current publishing model. As such, the reactions to it tend towards the extreme. Proponents herald the increased access to information, while critics decry the extralegal methods used to provide this access. In response to Elsevier's successful lawsuit (Association of American Publishers, 2017), John Willinsky, founder of the Public Knowledge Project, argues that although Sci-Hub was found guilty, "the academic community would likely view lack of access to journal articles as more serious for higher education than Elsevier losing money." (Elmes, 2017, para. 3). He also suggested the lawsuit indicates Elsevier's intention to turn research, frequently funded with public monies, into private property and corporate assets. The implication of Willinsky's statement is that while Sci-Hub's actions may be illegal, they are not immoral or they are at least less immoral than Elsevier's push to turn public goods into private profits.

Conversely, Maria Pallente, President and CEO of the Association of American Publishers, stated "'As the final judgment shows, the Court has not mistaken illegal activity for a public good. On the contrary, it has recognized the defendants' operation for the flagrant and sweeping infringement that it really is and affirmed the critical role of copyright law in furthering scientific research and the public interest'" (Association of American Publishers, 2017). Here Pallente conflates legality with morality as she attempts to frame the multi-billion-dollar industry she represents as the victim in this scenario.

The American Chemical Society (ACS) is, per volume of output, one of the most downloaded publishers in Sci-Hub (G. J. Gardner, McLaughlin, & Asher, 2017). In 2017, ACS filed a suit in Virginia similar to Elsevier's, demanding Sci-Hub cease distributing ACS content and asking for $4.8 million in damages. The ACS suit also seeks to have search engines, Internet service providers, and domain name registrars cease facilitating access to Sci-Hub (Kwon, 2017). The ACS suit represents an attempt to use state power to censor the Internet as a way of tamping down a challenge to the publishers' dominance of the conversation surrounding access to information.

From the publisher perspective, academic literature is a commodity, no different than any other consumer product. This sentiment is best expressed by Alicia Wise, director of universal access at Elsevier, in a New York Times article, "'It's as if somehow stealing content is justifiable if it's seen as expensive, and I find that surprising. It's not as if you'd walk into a grocery store and feel vindicated about stealing an organic chocolate bar as long as you left the Kit Kat bar on the shelf.'" (Murphy, 2016). The insinuation is that publishers have procured the materials for this good, then processed it for sale and consumption. One fallacy in this analogy is that unlike the cocoa beans, sugar, and other ingredients required to produce chocolate, academic literature is produced by scholars and given away for free for the advancement of science and the betterment of society. The second flaw in the analogy, is that unlike an overpriced candy bar, unaffordable academic literature still likely contains materials that are needed (J. Peters, 2016). The unintended outcome of Wise's analogy is a near-perfect example of the commodification of information, divorced from its production and creators.

As for the creators of academic literature, the *Chronicle of Higher Education* reached out to the authors of some of the most-downloaded works. The responses ranged from support for access with some trepidation over the long-term ramifications for society journals to considering it an honor (Ruff, 2016). In a survey of academics, 88% of respondents said that it was not wrong to download pirated papers, including 84% of respondents who had not used Sci-Hub. Additionally, 79% of respondents over age 50 had no problem with using Sci-Hub, so it is not simply a matter of generational divide. While over 50% of respondents use Sci-Hub or another method of guerilla open access to get articles they don't have access to, 17% do so because it is more convenient, and nearly a quarter of respondents use it because they object to the profits of commercial publishers (Travis, 2016). While neither of these pieces utilizes a rigorous methodology, they do provide a general insight into the views of both the creators and readers of academic literature. The only party that seems opposed to using Sci-Hub is the middle man who profits from the existing model, the commercial publishers.

Publishers contend that while they do not pay authors or reviewers, “they help ensure accuracy, consistency, and clarity in scientific communication” (McNutt, 2016). They also check quality, create visualizations, and promote content to media outlets. Publishers develop the online platforms that host the content. Publishers establish brands and cultivate and maintain good reputations. They manage the peer review process (Anderson, 2016; McNutt, 2016). While many of these tasks are necessary and Sci-Hub does not provide a mechanism for most of them, none of these tasks, or the remainder of the 96-item list of publisher actions listed by Anderson (2016), require a profit margin.

Several opinion pieces have come out against Sci-Hub. One of these compares Sci-Hub to mob accountant Meyer Lansky, mocks the struggles of researchers needing access to articles, and laments that publishers haven’t found an alternative to PDFs that would make unauthorized access more difficult (Esposito, 2016). Another piece parrots the lines about free or low-cost access alternatives like interlibrary loan, Research4Life, and university repositories while belittling Elbakyan, saying “She sincerely believes that she is above the law” (Cochran, 2016, para. 2). It should also be noted that neither of the aforementioned opinion pieces were written by faculty, but instead come from a contractor and a publishing company employee.

Not all opinion pieces on Sci-Hub are so negative or patronizing, some veer in the complete opposite direction. Oxenham (2016) and Heathers (2016) both refer to Elbakyan as the

Robin Hood of science. While these pieces can be a bit over-the-top, they do highlight the praxis behind Sci-Hub. The system, as both authors point out, is easier to use than the information silos provided by the publishers and it matches people with the information they need.

Falling in the middle between fawning and derision, some views highlight the problematic nature of a piracy system while highlighting the previously unmet needs that Sci-Hub addresses. In an interview with NPR, a SPARC spokesperson asks the essential question, “should such businesses be built around information that’s vital to the public’s good and the public’s health?” (NPR, 2016, para. 5). She goes on to explain that researchers generally have legal access to 50-70% of the literature they need for their work and that the current model is untenable and forces researchers to use systems like Sci-Hub to complete their work (NPR, 2016). While SPARC still continues to advocate for legal open access (SPARC, n.d.), the interview elucidates that access is a greater priority than legality for the group. The contradiction inherent in access limitations and the information needs of the people is an example of how the techno-capitalist nature of the existing system perpetuates inequalities.

The divide in how Sci-Hub is viewed seems to be based on people’s relationship to academic literature. For those who profit from its commodification, or are associated with those profiteers, Sci-Hub is viewed as a threat, a nuisance, and a criminally bad actor (Association of American Publishers, 2017; Cochran, 2016; Esposito, 2016; McNutt, 2016; Murphy, 2016). For those who create and use academic literature, Sci-Hub is viewed more as an equalizing force, at best, and an unfortunately necessary solution to a greater problem, at worst (Elbakyan, 2015; Elmes, 2017; NPR, 2016; Ruff, 2016; Travis, 2016).

2.2.2 Sci-Hub Usage

Since its founding in 2011, there have been three data releases regarding usage. The first, in early 2014, related to LibGen, or Library Genesis, the aforementioned backend storage for the Sci-Hub front end search and retrieval system (Cabanac, 2016). The second data release, in 2016, is the one used in this study and contains information related to individual download requests, specifically the date and time of the request, the digital object identifier (DOI) of the article requested, and the geographic coordinates of the nearest city to the request (Bohannon & Elbakyan, 2016). Lastly, the third release, in 2017, contains the list of 62 million DOI for the content stored by Sci-Hub (Greshake, 2017b; Hahnel, 2017).

In addition to scientific articles, LibGen contains scientific books and textbooks, along with some fiction books and comics. According to Cabanac (2016), in January 2014, LibGen contained nearly 23 million scientific articles equaling 15 terabytes of data. In addition, it contained over 1.1 million scientific books and textbooks equaling 13 terabytes of data. Beginning in late October 2012, articles were collected with a median 2720 articles added each day; however, the mean articles cached per day is nearly 53,000 due to 13 different days where more than 100,000 articles were added. Cabanac further analyzed the data and found that LibGen contained at least one article from 78% of journals published by DOI registrants, including 64% of Elsevier journals, 53% of Springer, and 59% of Wiley. While these major publishers have lower percentages of journals covered than the overall rate, they were far outpaced in terms of the percentage of papers covered. The overall average was just 36% of articles registered with a DOI, but the rates for these publishers were 77%, 53%, and 73% for Elsevier, Springer, and Wiley, respectively. Lastly, clinical medicine and chemistry were the fields with the most articles available. While not an examination of the Sci-Hub platform, this examination of the data demonstrates an overrepresentation of the major publishers, which is not surprising considering their combined share of the academic literature marketplace.

In 2016, *Science* correspondent John Bohannon reached out to Sci-Hub creator Alexandra Elbakyan in hopes of getting access to Sci-Hub usage data. The resultant data (Bohannon & Elbakyan, 2016) contains every download event from September 2015 through February 2016; each record contains the date and time of the request, along with the nearest city and the article's DOI. The results of Bohannon's (2016b) analysis show that Sci-Hub is being used all over the world, both in developing nations without much access to academic literature and from the wealthiest nations with, presumably, much greater access. Nearly 25% of Sci-Hub usage during this period came from the 34 member nations of Organization for Economic Cooperation and Development, the world's wealthiest nations; the United States was the 5th largest downloader, behind Iran, China, India, and Russia (Bohannon & Elbakyan, 2016). Bohannon's analysis also shows 3 million unique IPs for users; however, the real number of users is likely higher as many university users can share a single IP and Iran has downloaded a great deal of Sci-Hub data and created local, mirrored sites that provide Sci-Hub's content from Iranian servers instead of going directly to Sci-Hub's website (2016b).

The 2016 data release also contains the DOIs of the downloaded articles. Of the approximately 28 million download events, over 9 million came from Elsevier journals, followed by 2.6 million from Springer, 2.1 million from the Institute of Electrical and Electronics Engineers, nearly 1.9 million from the American Chemical Society, 1.3 million from Wiley, and 1.1 million from Nature (Bohannon, 2016b). The disproportionate downloads from Elsevier journals, comprising nearly 1/3 of all downloads, makes it unsurprising that Elsevier is Sci-Hub's most vocal and litigious opponent.

Greshake (2016) examined the download request data to better understand whether downloads were made by the general public or by academic researchers who couldn't get subscription-based access to the articles. Looking at the days and times of the downloads, he found that the heaviest use times were between 9am and 5pm. He asserts the data suggests that since most downloads occur during work hours, Sci-Hub is not simply used after work hours by the general public or academics from home. While Greshake suggests this is likely academics accessing journals that their institutions do not subscribe to, the analysis of the time of day download requests are made is not divided by country. While this may be more true for academics in developing nations, it also includes academics in wealthier nations where they are assumed to have access to their universities subscription-based journals and could support the publisher's argument that Sci-Hub usage is for convenience, not necessity (Bohannon, 2016b). The counterargument to this view, which would support Greshake's assertions, is access to licensed materials on-campus is frequently based on the user's IP address, meaning there would be no need to use a third-party service like Sci-Hub. Going through Sci-Hub on campus would create an unnecessary extra step that runs counter to PLE.

In addition to looking at the days and times of Sci-Hub downloads, Greshake (2016) also compared Sci-Hub usage with data from the World Bank. He found positive correlations between Sci-Hub usage and the population, Gross Domestic Product, Internet usage and availability, and life expectancy of countries. National unemployment was found to be a poor predictor of Sci-Hub usage. These correlations suggest that larger, wealthier countries with greater Internet infrastructure are more likely to use Sci-Hub.

Finally, in order to better understand academics' Sci-Hub usage, Greshake (2016) used a list of global college and university IP ranges to compare with Sci-Hub data. Greshake discloses that this IP list is both slightly outdated, 18 months old at the time of the study, and incomplete; these caveats, he argues, suggest that the findings be taken as low estimates. While the Sci-Hub data was scrubbed of IPs due to privacy concerns, Greshake reached out and asked for the data and Sci-Hub returned the data in two forms. The first dataset shows that on-campus Sci-Hub usage tends to be between 8 and 10 percent on workdays and drops during weekends and over holidays. The second dataset breaks down the data by country, but limits it to 10 day periods instead of 24 hours to help preserve anonymity. The findings from this dataset were a bit more suspect as the small number of downloads in some countries may have skewed the results and a possible lack of academic IP ranges for other countries may underestimate academic usage. Overall, the findings suggest little correlation between the percentage of population enrolled in higher education and Sci-Hub usage and the relationship between national research funding and Sci-Hub usage was also small. While interesting, Greshake's use of global data may make understanding any one nation's Sci-Hub usage more difficult. The findings do suggest academic use while on campus is relatively high and provides insight into how the global higher education landscape uses Sci-Hub.

While admittedly meant to be a rough estimate, Bianca Kramer (2016a, 2016b) of the Utrecht University Library built upon the work of Greshake (2016) and Bohannon (2016b) to examine access versus convenience for the Sci-Hub download requests attributed to her university and the Netherlands more generally. For part one of her analysis, Kramer (2016a) limited the dataset to Sci-Hub requests to the Netherlands. The dataset was then split by cities with 1000 or more Sci-Hub download requests; Amsterdam was removed from the list because many Dutch internet service providers are located there and the actual user could be from anywhere in the Netherlands. The results of the analysis showed that both in raw numbers and when compared to population size, towns and cities with universities, in general, had more downloads. However, Den Haag, the home to the Dutch parliament and many international organizations, had more download requests than all but two other cities despite not having a university. The author suggests looking at cities with research companies could help clarify Sci-Hub usage, especially from non-university cities. Overall, this analysis suggests that cities with universities may be more likely to use Sci-Hub, but it is clearly not only campus communities who are using the site.

In the second part of the study, Kramer (2016b) found nearly 3000 unique DOIs attributed to Utrecht University in Greshake's (2016) study and manually checked each DOI to see if the article was available either through library subscription or because it was available via open access or otherwise freely available. The overall findings showed that 60% of the articles were available through subscriptions, 15% were available for free, and 25% were not available via the publisher. Interestingly, it is clear that the university does not subscribe to many journals of certain publishers as 98% of articles published by IEEE were not available, nor were 78% of De Gruyter, 55% of IOP Publishing, 65% of American Institute of Physics, and 86% of Cambridge University Press. While the author looks at open access journal availability, she admittedly does not attempt to examine access through pre-print servers or institutional repositories (Kramer, 2016b). At minimum, this study implies that a quarter of the Sci-Hub usage attributed to Utrecht University was based on need, not convenience.

To examine Sci-Hub usage in Latin America, Machin-Mastromatteo, Uribe-Tirado, and Romero-Ortiz (2016) replicated Bohannon's (2016b) analysis, but limited the results to Latin America, which included 32 countries for the purpose of this study. Latin America represents approximately 12.5% of the worldwide download requests, which surprised the authors due to the limited funds most Latin American countries have to spend on journal subscriptions. Brazil represents nearly 30% of the Latin American usage, followed by Mexico, Chile, Colombia, Argentina, and Peru; combined these six nations represent almost 90% of all Latin American Sci-Hub download requests. The authors added no new methods of analysis to Bohannon's study, nor do they compare Sci-Hub downloads to population size for the represented countries. Gardner, McLaughlin, and Asher (2017) produced a wide-ranging study on how Sci-Hub download requests are related to aspects of the academic literature marketplace. As outlined in chapter one, interlibrary loan (ILL) is the method used by libraries to share books and articles that a library has with another library without access to that material. As Sci-Hub can potentially eliminate, or at least minimize, the access limitations that prompt ILL usage for articles, Gardner et al. (2017) use the 2016 Sci-Hub data release to examine how ILL has been affected by Sci-Hub in the United States and Canada and how Sci-Hub downloads compare to licit downloads. The ILL portion of the study was conducted using multiple methods. The first method involved selecting large higher education institutions that were the biggest institutions in their municipality and where the municipality did not have multiple large colleges or universities; the authors reached out to dozens of institutions that met these criteria, with ten universities agreeing to participate. The ten universities sent ILL requests for items that were not full books or media for the same time span as the Sci-Hub data, September 2015 through February 2016. The authors studied the effects of Sci-Hub usage on ILL by using repeated measures analysis of variance (rANOVA) and multivariate analysis of variance (MANOVA). None of the cities showed a significant negative correlation between ILL demand and Sci-Hub downloads.

Gardner et al. (2017) also used the Sci-Hub data to compare with data from the Association of College and Research Libraries (ACRL). Unfortunately, the most recent ACRL dataset available at the time, from 2014-15 academic year, does not overlap with the Sci-Hub data timeframe. The authors still performed a geographical analysis on the most recent ACRL data and found that Sci-Hub downloads are moderately and positively correlated with the number of academic institutions, total library expenditures, total ILL articles borrowed, full-time instructional faculty, and full-time graduate students; total ILL articles borrowed is more strongly correlated with each of these variables. While the findings are interesting and some of the methodologies match the methodologies in this study, the mismatched years make it difficult to ascertain the importance of the results.

Additionally, Gardner et al. (2017) examined whether institutional subscription price or findability via abstracting and indexing databases made a journal more or less likely to be downloaded via Sci-Hub. The authors limited the Sci-Hub data to requests from the United States, then further limited it to titles classified by CrossRef as either a journal or journal article. A random sample of 270 titles, from the 44,068 entries, was generated to allow for a 90% confidence level with a 5% margin of error. The authors manually checked each entry against the Ulrichsweb Global Serials Directory and noted the price and the number of abstracting and indexing databases the entry is listed in. Subscription price was found to not be a monocausal explanation of Sci-Hub usage, but was correlated with download count; the Spearman value ρ=.602 led the authors to believe a power-law relationship may be in effect. Inclusion in abstracting and indexing services was found to not be a determining factor in Sci-Hub usage (G. J. Gardner et al., 2017). The simplicity of Sci-Hub browser extensions for Google Scholar may explain why inclusion in indexing and abstracting services does not play a determining role in Sci-Hub usage since Google Scholar does not rely on traditional library-centered services like those provided by EBSCO.

Lastly, comparing Sci-Hub downloads to licit downloads as a way to examine lost revenues for publishers, Gardner et al. (2017) found download estimates for four of the top ten downloaded publishers (Bohannon, 2016b). With only six months of Sci-Hub data available, Sci-Hub download data was doubled. Ratios of publisher to Sci-Hub downloads were generated; however, the authors note that since Sci-Hub data was collected during the academic year compared to the yearly totals for publisher downloads, the ratio should be considered a rough measure. The results show that Elsevier, the most illicitly downloaded publisher, had the greatest revenue lost, $334.21 million, when using the highest price per article purchase the publisher lists. The Royal Society of Chemistry had the greatest proportional loss due to the lowest ratio of licit to illicit downloads, 20:1, meaning that for every licit download, ACS had 20 illicit downloads. While their study does not necessarily answer any of the research questions posed in this study, the methodologies provide interesting approaches for how to study the Sci-Hub data, specifically using large institutions that are the only large institutions in their respective cities.

With the release of 2017's Sci-Hub data (Hahnel, 2017), there is now a dataset containing a list of every article in the Sci-Hub/LigGen repository. The first number that draws attention is the sharp increase in the volume of articles from the 23 million at the beginning of 2014 (Cabanac, 2016) to nearly 63 million in March 2017 (Hahnel, 2017), a difference of approximately 40 million articles. Greshake (2017a) combined this new dataset with the 2016 data (Bohannon & Elbakyan, 2016) to match on the article DOI and provide insight into how often each article in Sci-Hub was requested. His findings show that despite articles going back as far as 1619, 95% of articles downloaded were published after 1982 and 35% were less than 2 years old at the time of download.

While over 177,000 journals are represented within Sci-Hub, less than 10% of the journals comprise over 50% of the total content. Furthermore, less than 1% of the journals are getting over 50% of the total downloads. Similarly, while there are approximately 1700 publishers within Sci-Hub, the top 9 publishers comprise roughly 70% of the total articles and 80% of all downloads (Greshake, 2017a), these findings align with Zipf's Law (1949) as it has been applied to information seeking, where it is has been called the 80/20 or 70/30 rule (Case & Given, 2016) and applied to article requests (Dorsch & Pifalo, 1997) and citations (White, 2001). Of the most frequently downloaded journals, chemistry appears to be the most popular topic with chemistry journals representing 12 of the top 20 most downloaded titles. The author posits this could be a result of undersupplying by university libraries or that chemistry and engineering graduates are more likely to go into private, for-profit industries than their peers in the health sciences who tend to stay within academic institutions (Greshake, 2017a). This analysis shows how a small number of publishers dominate the most requested articles and how different fields use Sci-Hub in varying ways.

Himmelstein, Romero, McLaughlin, Tzovaras, and Greene (2017) offer additional analysis of the 2017 dataset. For their study, all Sci-Hub entries are referred to as articles, although the actual entries consist of conference proceedings articles, journal articles, book chapters, and reference entities, amongst other things. A primary focus of their study was assessing coverage or examining the percentage of articles in Sci-Hub versus the total number of articles in existence. By subject area, chemistry and chemical engineering had the highest coverage levels with both at 92% or higher; conversely, arts and humanities, multidisciplinary subjects, and computer science had less than 78% coverage. The high coverage levels for chemistry fields is not surprising based on Greshake's (2017a) findings that chemistry was the most popular topic.

Taken as a whole, the body of literature on Sci-Hub usage suggests that it varies widely by field and by location and conclusive arguments cannot yet be made. The popularity and coverage of fields such as chemistry (Greshake, 2017a; Himmelstein et al., 2017), can be interpreted as supporting the claim that Sci-Hub is used by those outside academia since those with advanced degrees in the field frequently go to work in the private sector (Greshake, 2017a). Kramer's (2016a) finding of high Sci-Hub usage in Den Haag, a city with the Dutch parliament and international organizations, but no universities, also aligns with this idea. While, Greshake's (2016) finding that most Sci-Hub use comes during traditional work hours and he asserts this as academics using the system while on campus, it could also be researchers outside of higher education. However, Sci-Hub usage tends to be highest in cities with large universities (Bohannon, 2016b; Kramer, 2016a), so while the existing studies provide some insight, especially globally, there are still many unanswered questions, including whether Sci-Hub is being used to access otherwise inaccessible materials or if it is being used because it requires the least effort to meet information needs.

2.3 LIMTATIONS AND GAPS IN THE LITERATURE

As Sci-Hub was created in 2011 and its actions are legally questionable, it is unsurprising to find a limited number of studies related to how and why it is used. However, the volume of Sci-Hub related studies is growing rapidly. Most examinations of the existing Sci-Hub data focus on global usage (Bohannon, 2016b; Cabanac, 2016; Greshake, 2016, 2017a; Hahnel, 2017;

Himmelstein et al., 2017); this trend makes sense as Sci-Hub was created by a researcher in a country with limited access to academic literature and was designed to help those in similar situations (Bohannon, 2016a). However, many of the legal troubles facing Sci-Hub stem from the United States and focus on how the system is used within the United States (Kwon, 2017; United States District Court Southern District Of New York, 2015).

Global disparities in access to academic literature are well-recognized and acknowledged even by publishers. To address some of these gaps, the United Nations has worked with publishers on a number of programs through the Research4Life initiative ("Research4Life home," n.d.). Beyond problems such as withdrawing access to journals in certain countries (K. Kmietowicz et al., 2011; Z. Kmietowicz, 2011), these programs also only cover a limited range of fields. Considering the great variation in access between countries in the developing world and those in more industrialized nations, global overviews of Sci-Hub usage are limited in how they can explain usage in any one country.

With the exception of Greshake's (2016) study comparing Sci-Hub usage with World Bank indicators and known college and university IP ranges and Gardner, McLaughlin, and Asher's (2017) study of Sci-Hub's effect on ILL, previous studies primarily report on descriptive statistics and do not delve into any relationships that might provide a greater depth of understanding. As Greshake notes in his work, there are numerous possible problems with his study, but it does provide a good first step toward understanding how Sci-Hub usage is related to other factors. The other studies (Bohannon, 2016b; Cabanac, 2016; Greshake, 2017a; Hahnel, 2017; Himmelstein et al., 2017) help provide context for what is used on Sci-Hub overall and how frequently it used, but there is still a need for greater depth even at the expense of breadth.

This study was intended to help fill some of the existing gaps in literature. By focusing on only the United States and forgoing a larger global context, the hope is to provide a greater granularity of understanding. Bohannon (2016b) notes that many Sci-Hub download requests came from New York City, home to many universities and scientific institutions, and also smaller cities like Columbus, Ohio and East Lansing, Michigan with large universities. The implication is that it is members of these campus communities that are using Sci-Hub, which would support publisher claims of convenience over necessity. However, neither Ohio State University or Michigan State University, located in Columbus and East Lansing, respectively, subscribe to all academic journals; even resource-rich Harvard has had to make cuts to their journal collection (Sample, 2012). There has previously been no study showing whether the download requests are for journals these institutions subscribe to or not. One of the goals of this study was to help provide a better understanding of that for the United States, similarly to what Kramer (2016b) did on a smaller scale for Utrecht University, albeit without using IPs specifically tied to higher education institutions.

Additionally, where Greshake (2016) compared Sci-Hub download request data with known college and university IPs to try to understand how many people are using Sci-Hub while on campuses, this study looked at how download requests on or near a research university campus are related to the academic journal expenditures and other characteristics of that institution. Greshake drew on his own experiences of not having access to all pertinent literature at the University of Frankfurt to develop his line of inquiry and this study did something similar but limits the scope of the study to the United States instead of using global data.

Greshake (2016) also correlated the Sci-Hub download request data with World Bank indicators, including enrollment in higher education and research expenditures. This study took a similar approach for the United States by breaking the country down to smaller geographic regions and looking at the percentage of residents with advanced degrees while accounting for the higher education institutions in the region as well as the number of faculty and graduate students. The first part of the study also used research funding by institution as a predictor variable for Sci-Hub usage. The second part of the study looked at how the presence of colleges and universities in a region, along with the concentration of faculty, graduate students, and advance degree holders, are related to Sci-Hub usage, similar to Kramer's (2016a) comparison of cities with universities versus non-university cities in the Netherlands.

The relative newness of Sci-Hub and the even newer releases of data creates an environment ripe for exploration. This study helps to provide a greater understanding of who in the United States is using Sci-Hub and the possible motivations. While this study cannot definitively ascribe motivations to Sci-Hub users, looking at the relationship between the number of Sci-Hub download requests and the characteristics of the academic institutions and the people near those requests provides some much-needed insight into this phenomenon.

3.0 METHODOLOGY

The purpose of this study was to examine how Sci-Hub usage, as measured by the number and location of download requests, relates to the environment surrounding the download requests. Sci-Hub critics, especially those with connections to the for-profit academic publishing industry, suggest that Sci-Hub usage in the United States is based on convenience, not necessity, as there are alternative methods in place to provide access to paywalled content for those without subscriptions (Bohannon, 2016b; McNutt, 2016). The primary subscribers to academic journals are academic libraries and the clusters of Sci-Hub requests near higher education institutions lends some credence to this theory (Bohannon, 2016b); however, there has been no previous analysis that examines the relationship between Sci-Hub usage in the United States and other possible contributing factors. This study is broken down in two parts: 1) how Sci-Hub download requests are related to the institutional characteristics of research-intensive universities and 2) how Sci-Hub download requests are related to the population of their geographic location.

3.1 RESEARCH QUESTIONS

3.1.1 Sci-Hub and Academic Libraries

1. How do the academic journal expenditures of research institutions relate to the number of Sci-Hub article requests within a 10-mile radius of the institution?

    1.1. For Sci-Hub article requests within a 10-mile radius of a research institution, are the requests for articles that appear in journals for which the institution has a subscription?

3.1.2 Sci-Hub and Population

2. What is the relationship between the percentage of residents over 25 with an advanced degree in a core-based statistical area (CBSA) and the number of Sci-Hub requests?

    2.1. How does the number and type of higher education institutions and the total number of graduate students within a CBSA change the above relationship?

3. What is the relationship between the percentage of residents over 25 (minus the total faculty within the region) with an advanced degree in a CBSA and the number of Sci-Hub requests?

    3.1. How does the number and type of higher education institutions and the total number of graduate students within a CBSA change the above relationship?

3.2 DATA SOURCES

3.2.1 Sci-Hub

In 2016, John Bohannon (2016), a contributor at *Science*, noticed that despite several opinion pieces on Sci-Hub, there was very little information about who uses the system, where the users are located, and what articles are being downloaded. To help rectify the situation, Bohannon reached out to Sci-Hub creator Alexandra Elbakyan. Over the course of several weeks, the two collaborated to create a public dataset that would provide the DOI of every article requested, along with non-identifying location information. The Sci-Hub data comes from the Sci-Hub server logs for request events from September 1, 2015 through February 29, 2016 (Bohannon, 2016b; Bohannon & Elbakyan, 2016). Every download event is included in the original dataset, including the date and time of the request, with the exception of eight days in November when Sci-Hub underwent a domain change due to a lawsuit by Elsevier. Using data from Google Maps, users' geographic locations were aggregated to the nearest city to protect their privacy (Bohannon, 2016b); the fields for the geographic locations consist of the city name, country, and geographic coordinates. Additionally, the dataset contains the date and time of the request. This study included only results in the United States; the full set of Sci-Hub requests from within the United States has 1,150,963 document download requests.

3.2.2 IPEDS

Institutional data comes from the Integrated Postsecondary Education Data System's (IPEDS); all institutions that participate in any federal student financial aid program are required to complete all IPEDS surveys. The 2015 IPEDS dataset was selected to match the timeframe of the Sci-Hub data. Specifically, this study includes data from the Institutional Characteristics File, the Academic Libraries File, the Employees by Assigned Position File, and the Fall Enrollment File. Eliminating institutions whose Carnegie classification did not include offering at least an Associate's degree (see Appendix A) left 3,183 higher education institutions in the United States. The Institutional Characteristics File contains the directory information of every institution contained in IPEDS. Generally, this data source contains information on if an institution is degree-granting and, if so what the highest degree offered is, the Carnegie Classification, whether the institution is public or private, and other basic categorical information. Specifically, the Institutional Characteristics File includes information on institutional type, size, geographic location (including coordinates and CBSA), and whether the institution has a medical school or a hospital (see Table 1).

The Academic Libraries File includes characteristics of the library, such as the number of books and other materials, circulation data, number of branches, staff size, interlibrary loan data, and expenditures. For the purpose of this study, the most important field in the Academic Libraries File is yearly academic journal expenditures (see Table 1).

The Employees by Assigned Position File contains the number of staff, classified by full-time and part-time status. Included in this file is whether an employee is part of a medical school, tenure status, and occupational category. For the purpose of this study, the most important field in the Employees by Assigned Position File is the number of faculty at an institution (see Table 1).

Lastly, the Fall Enrollment File contains the total number of enrolled undergraduate and graduate students; these fields are further broken down by race, gender, and nonresident alien status. For the purpose of this study, the most important field in the Fall Enrollment File is the number of graduate students (see Table 1).

3.2.3 Census

The Educational Attainment data comes from the U.S. Census Bureau, American Community Survey 1-Year Estimates (US Census Bureau, n.d.-b). The American Community Survey (ACS) uses core based statistical areas (CBSAs) to group together geographic regions that contain at least one urbanized area or urban cluster of at least 10,000 people plus the surrounding areas that have a high degree of economic and social integration (US Census Bureau, n.d.-a). There are 546 CBSAs in the dataset. This dataset includes the education level, by total number and percentage, of a CBSA's population. The education levels are less than high school graduate, high school graduate, some college or associate's degree, Bachelor's degree or higher, and graduate or professional degree. The data is further dissected by age ranges: 18 to 24 years, 25 to 34 years, 35 to 44 years, 45 to 64 years, and 65 years and older.

Table 1. Descriptive Statistics

Variable	Source	Mean	Min	Max	STD
Journal Expenditures (R1 & R2)	IPEDS	6419360	347601	20,400,000	3883662
Graduate Students (R1 & R2)	IPEDS	7246.46	210	27569	4816.61
Faculty (R1 & R2)	IPEDS	2019.60	84	9362	1443.34
Medical School (R1 & R2)	IPEDS	1.52	1	2	0.50
Hospital (R1 & R2)	IPEDS	1.81	-1	2	0.43
NIH (R1 & R2)	NIH	27,600,000	0	453,000,000	70,300,000
NSF (R1 & R2)	NSF	150,118.60	0	1,369,278	248,647.20
CBSA Population	Census	379,020.90	36777	13,900,000	987,025.30
CBSA Population w/ Adv Degree	Census	46,415.49	1483	2,197,544	145,055.50
Faculty in CBSA	IPEDS	2415.61	0	86,665	6067.29
Graduate Students in CBSA	IPEDS	6757.56	0	234,513	19,414.76
Associate’s in CBSA	IPEDS	1.96	0	37	3.50
Bachelor’s in CBSA	IPEDS	1.32	0	25	2.53
Master’s in CBSA	IPEDS	1.28	0	44	2.99
Doctorate in CBSA	IPEDS	0.63	0	20	1.63

For the purpose of this study, this dataset provides the population over 25-years-of-age in a CBSA and the population over 25-years-of-age with advanced degrees in a CBSA; advanced degree holders are defined as having a graduate or professional degree. The 2015 estimates were used to match the timeframe of the Sci-Hub data. The 2010 ZIP Code Tabulation Area to Metropolitan and Micropolitan Statistical Areas Relationship File (US Census Bureau, n.d.) was used to match the computed ZIP codes of Sci-Hub requests with their corresponding CBSAs. The CBSAs are calculated prior to each census; the 2010 calculations are the most current and applicable to the 2015 American Community Survey.

3.2.4 NIH and NSF

As two of the largest funders of research at higher education institutions, the National Institutes of Health (NIH) and the National Science Foundation (NSF) grant expenditures were selected as representatives of research performed at an institution. The 2015 NIH awards data (NIH, n.d.-a) was selected to match the timeframe of the Sci-Hub data and matched against all 222 Research I and Research II universities. The dataset includes the total grant expenditures per institution, the city and state of the institution, and the total number of awards that institution received; the total grant expenditures per institution (see Table 1) is the field used in this study.

Additionally, the NSF research and development expenditures (NSF, n.d.) was selected as a second representative of research at an institution. The 2015 NSF data was selected to match the timeframe of the Sci-Hub data and matched against all 222 Research I and Research II universities. The NSF dataset provides the total expenditures per institution and the rank and percentile of each institution; the total expenditures per institution (see Table 1) is the field used in this study. Both datasets are publicly available as part of institutional reporting requirements for transparency.

3.2.5 Data Summary

As outlined above, this study is broken into two parts, the relationship between Sci-Hub download requests and the characteristics of nearby research institutions and the relationship between Sci-Hub download requests and the population and characteristics of the geographic area where the requests are generated from. In the first section, this study isolated Sci-Hub download requests near research institutions and examined how journal expenditures and other institutional characteristics are related to the number of Sci-Hub download requests. In the second section, this study examined the relationship between Sci-Hub download requests and the percentage of the nearby population with advanced degrees. This section further explored how that relationship changes when the number and type of higher education institutions, along with the graduate students and faculty of these institutions, were accounted for.

3.2.6 Sci-Hub and Academic Libraries: Q1

3.2.6.1 Sample

Use of academic journal articles is typically highest amongst those conducting research, so the list of higher education institutions has been limited to the Carnegie classifications of “Doctoral Universities – Highest research activity” and “Doctoral Universities – Higher research activity,” known respectively as R1 and R2 institutions (“Basic Classification,” n.d.). Doctoral universities are divided between three categories: “highest research activity,” “higher research activity," and "moderate research activity" ("Basic Classification," n.d.). The categories are determined by research and development (R&D) expenditures in science and engineering, non-science and engineering R&D expenditures, and doctoral conferrals. Research activity is placed on two indices: an aggregate level of research activity and per-capita research activity. Institutions are evaluated based on a distance from a common reference point. Institutions that place very high on either index are assigned to the "highest research activity" group, institutions high on at least one, but not very high on either, are assigned to the "higher research activity" group, and the remainder are placed in the "moderate research activity" group ("Basic Classification Methodology," n.d.). The "moderate research activity" group, also known as R3 institutions, were omitted from the study due to their comparatively lower level levels of research and doctoral degree conferrals. Institutions that did not have a Carnegie classification of 15 or 16 (see Appendix A) were dropped from the dataset.

The total number of R1 and R2 institutions, based on IPEDS data, is 222. The sample for this section was further limited by geographic proximity to other R1 and R2 universities. Isolated research institutions (IRI) were determined by using the geonear (Picard, n.d.) command in Stata that measures the shortest distance between one set of geographic coordinates and another set of coordinates. Using this procedure, the distance to the closest R1 and R2 institutions for each of the 222 universities was calculated. The distance of 20 miles was chosen to provide the most inclusive number of institutions while attempting to avoid possible overlaps; any R1 or R2 institution with a calculated distance of 20 miles or less was dropped from the dataset. This proces lowered the number of institutions in this part of the study to 136. The 20-mile distance ensured that, for the purpose of this study, all Sci-Hub requests within a 10-mile radius of an institution were associated with that institution and could not be associated with any other IRI.

The Sci-Hub download requests were narrowed by limiting the Sci-Hub download requests to only those within 10 miles of an IRI. Using the geonear (Picard, n.d.) command in Stata again, the coordinates of IRIs were compared to the coordinates of the Sci-Hub requests. All requests within 10 miles of an IRI were tallied using the contract command in Stata and associated with that institution. Any Sci-Hub download request made at a distance greater than 10 miles from an IRI was dropped from the dataset, resulting in the number of observations dropping from 1,150,963 to 419,934. For Q1, there are 136 Isolated Research Institutions and 419,934 Sci-Hub download requests. Figure 4 provides a visual representation of the IRI by the number of associated Sci-Hub download requests.

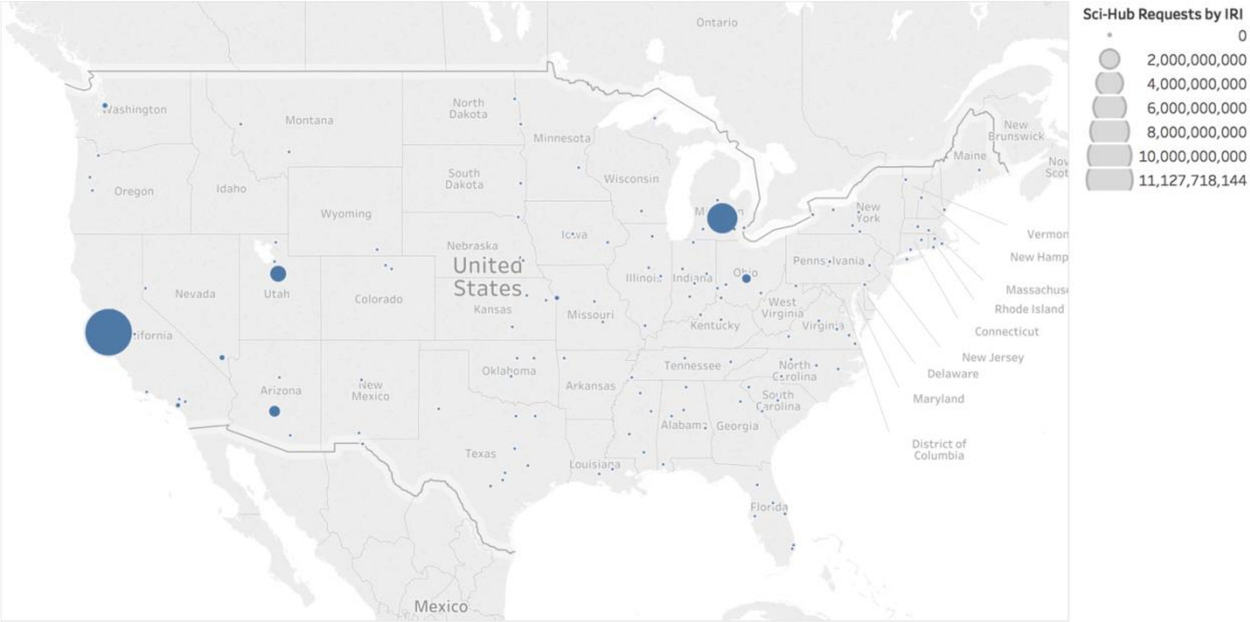

Figure 4. Isolated Research Institutions

This figure shows the 136 Isolated Research Institutions with the size representing the number of Sci-Hub download requests within 10 miles of campus.

The unique identifiers and institution names from the IPEDS dataset, limited to IRIs, was exported to a Microsoft Excel spreadsheet and research expenditures from NIH (NIH, n.d.-a) and NSF (NSF, n.d.) were manually added to the spreadsheet. The new research data was imported into the dataset using the unique identifiers as the matching variable. The NIH and NSF research funds were combined to form a new research variable; one institution did not receive money from either source. Additionally, the number of graduate students and faculty were combined to form a new graduate/faculty variable to represent institution size. Lastly, a new medical/hospital binary variable was created to represent if an institution had either a medical school or a hospital. To compensate for skewedness, the following variables were transformed using a base-2 logarithm: journal expenditures, Sci-Hub requests, research, and graduate/faculty.

Variables

Dependent Variable

The dependent variable for Q1, the number of Sci-Hub download requests, was determined by the number of Sci-Hub download request events within a 10-mile radius of an IRI. The distribution of the Sci-Hub requests is right-skewed as 90 IRIs have fewer than 1000 requests, but 7 IRIs have more than 10,000. As stated above, the number of Sci-Hub download requests has been transformed using a log base-2 algorithm to account for this skewedness. The resultant mean of the transformed number of Sci-Hub download requests is 9.26, with a minimum of 0 and a maximum of 16.69 (see Table 2).

Independent Variables

To better understand the relationship between the number of Sci-Hub download requests and academic institutions, several possible contributing factors were included as representatives of the need for access to academic literature and its availability: the size of an institution’s faculty and graduate student population, the prevalence of research funding, and the presence of medical training or practice. The independent variables for Q1 included academic library journal expenditures, the combined number of faculty and graduate students, research funds from NIH and NSF, and whether an institution has a medical school or associated hospital.

Academic library journal expenditures were used to represent the amount of academic literature that a campus community has access to (Tenopir, Volentine, & King, 2012). While subscription-based journals vary widely in costs, academic journal expenditures are the best available variable to represent access to paywalled academic literature. The distribution of academic journal expenditures is right-skewed; while 98 institutions spend less than $8,000,000 on academic journals, 22 spend more than $10,000,000 and 6 spend more than $15,000,000. The skewedness of the distribution required a log base-2 transformation. While the resultant distribution of the transformed variable is left-skewed, the skewness and kurtosis are closer to normal. The resultant mean of the transformed journal expenditures is 22.31, with a minimum of 18.41 and a maximum of 24.29 (see Table 2).

As academic journals are used for research purposes, along with graduate student instruction, graduate students and faculty are the most likely user base of academic literature (Abbott, 2016). The total number of graduate students and faculty at an institution were totaled to provide a representation of the academic journal user base. The distribution of graduate students and faculty is right-skewed; while the mean is approximately 8800 and the median is 7784, 6 institutions have more than 20,000. The skewedness of the distribution required a log base-2 transformation to achieve a more normal distribution. The resultant mean of the transformed total number of graduate students and faculty is 12.85, with a minimum of 9.58 and a maximum of 14.81 (see Table 2).

Performing literature reviews is a vital step in the research process. Institutions that perform more research may have a greater need for access to academic literature to pursue that research (Tenopir et al., 2012). The National Institutes of Health and the National Science Foundation are two of the largest funders of research; the 2015 funding data for both organizations were combined to create a general research funding variable to represent the amount of research performed at an institution. Like the previous variables in this model, research dollars have a right-skewed distribution; the skewedness is demonstrated by a $3.5 million difference between the mean and median values. The skewedness of the research funding distribution required a log base-2 transformation to provide a more normal distribution. The resultant mean of the transformed total research dollars from NIH and NSF is 23.61, with a minimum of 16.13 and a maximum of 28.76 (see Table 2).

Medical schools and hospitals have unique information requirements. In addition to the normal information needs of researchers, medical practitioners also require clinical information for the immediate diagnosis and treatment of patients. As such, institutions with either a medical school or an associated hospital may have increased expenditures on licensed information resources and a higher number of information requests. The presence of either a medical school or associated hospital was transformed into a binary variable to account for their uniqueness; 70 institutions have a medical school and/or an associated hospital while 66 do not.

Table 2. Q1 Variable Definitions and Descriptive Statistics, 136 Observations

Variable	Mean	Min	Max	STD	Measurement
Dependent Variable					
$\log_2$(Sci-Hub Requests)	9.26	0	16.69	2.49	Log base-2 of Sci-Hub requests within 10 miles of IRI
Independent Variables					
$\log_2$(Journals)	22.31	18.41	24.29	1.01	Log base-2 of academic journal expenditures by an IRI
$\log_2$(Grad/Faculty)	12.85	9.58	14.81	0.89	Log base-2 of the total number of graduate students and faculty at an IRI
Medical/Hospital	.51	0	1	0.50	0=No medical school or hospital; 1=IRI has medical school and/or hospital
$\log_2$(Research)	23.61	16.13	28.76	2.58	Log base-2 of combined research funds from NIH and NSF

Interaction Effects and Collinearity

The correlation matrix (Table 3) shows that each independent variable has a statistically significant correlation with the other independent variables. The high levels of correlation between independent variables is understandable as they seem to work in concert. A higher number of graduate students and faculty would suggest more individuals performing research and a wider range of disciplines which would require a wider range of journal subscriptions. Part of conducting research is performing literature reviews; greater research funding suggests a greater need for academic journals. The high correlation between the independent variables suggested the possibility of multicollinearity. Stata was used to determine the variance inflation factor (VIF). No VIF values greater than 10 were found, so there is likely no problem with multicollinearity (Acock, 2014). For the remaining variables, interaction tests were conducted on each pair of predictors. Interaction variables were generated by creating cross-products of the variables and inserting them into the model to determine interaction effects.

Table 3. Correlation Matrix

	Log$_2$(Sci-Hub)	Log$_2$(Journal)	Log$_2$(Grad/ Faculty)	Med School/ Hospital	Log$_2$(Research)
log$_2$(Sci-Hub)	1				
log$_2$(Journal)	0.39***	1			
log$_2$(Grad/ Faculty)	0.40***	0.71***	1		
Med School/ Hospital	-0.30***	-0.51**	-0.57**	1	
log$_2$(Research)	0.47***	0.65***	0.55***	-0.59***	1

3.2.7 Sci-Hub and Population: Q2 & Q3

3.2.7.1 Sample

The American Community Survey data consists of 546 entries; however, some of the CBSAs within the data are duplicated. For example, CBSA 10580 is listed twice, once as Troy, NY and once as Albany, NY; the population data for both city labels are the same. CBSA 14460 is listed five times for Chestnut Hill, Cambridge, Boston, Waltham, and Medford, MA. Duplicate entries have been dropped from the dataset; 508 CBSAs remained after deduplication removed 38 entries. Figure 5 shows a map of the United States with the CBSAs drawn; the Sci-Hub document download requests are overlaid on this map.

The IPEDS Institutional Characteristics File contains the CBSA and the Carnegie classification of each institution. The Employees by Assigned Position File was used to gather the total number of faculty at each institution and the Fall Enrollment File was used to gather the number of graduate students at each institution. The three IPEDS data files were merged on their shared unique identifier. Using the contract command in Stata, the total number of graduate students in each CBSA was obtained; the total number of faculty per CBSA was also obtained in this manner.

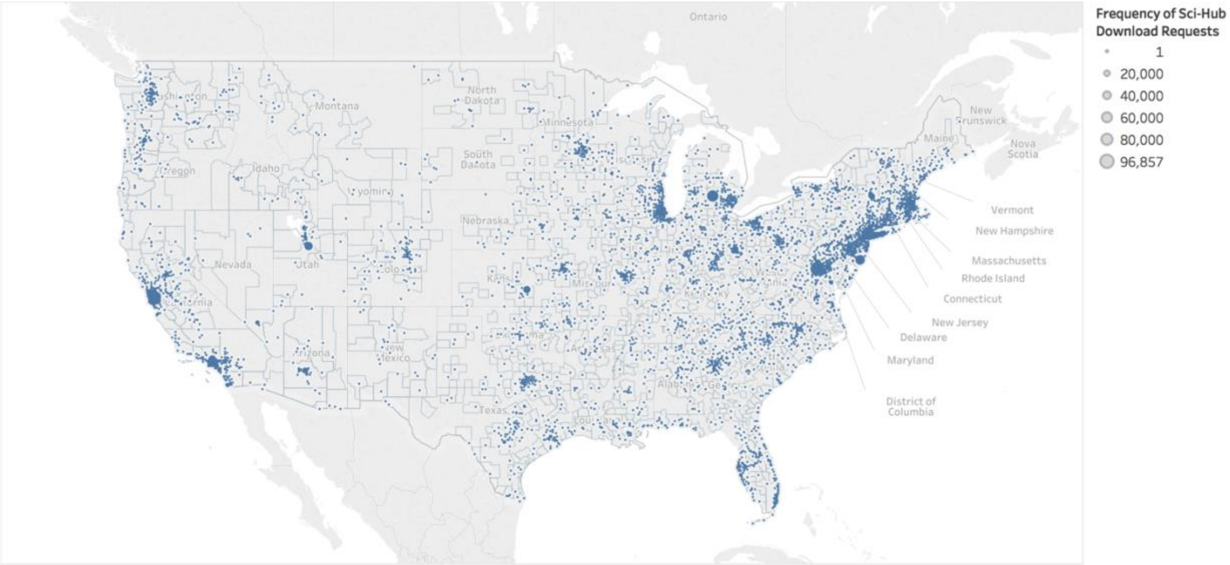

Figure 5. Sci-Hub Download Requests by CBSA

This figure shows Sci-Hub download requests overlaid on a U.S. map divided by CBSAs.

Variables

Dependent Variable

The dependent variable in Q2 and Q3, the number of Sci-Hub download requests, was determined by the number of article download requests made within a CBSA. The Sci-Hub dataset provides the geographic coordinates of the nearest town or city of each article download request. Using the opencagegeo command (Zeigermann, n.d.) in Stata, the coordinates of the download request underwent reverse geocoding through OpenCage Geocoder (OpenCage Data, n.d.) to obtain an address. Reverse geocoding takes the latitude and longitude of a point and translates them into a readable address. The 2015 HUD USPS ZIP Code Crosswalk file (U.S. Department of Housing and Urban Development, n.d.) is a table containing the ZIP codes within each CBSA. The resultant ZIP code after reverse geocoding was matched against the HUD Crosswalk file to obtain the CBSA. Using the contract command in Stata, the total number of

Sci-Hub requests per CBSA was tallied; this process counted the number of Sci-Hub download request entries for each CBSA code to create a total number per CBSA.

The HUD Crosswalk file uses a slightly different list of CBSAs than IPEDS, breaking up larger metropolitan areas into smaller subsections. This difference resulted in 0 requests for large cities like New York City, Los Angeles, and Washington; a manual examination of the data was performed to join the divided sections together to match the CBSAs in the IPEDS and Census data. This allowed for including major metropolises, which have already been shown to have large numbers of Sci-Hub download requests (Bohannon, 2016b). The distribution of Sci-Hub download requests is extremely right-skewed, with the mean of nearly 6000 and a median of only 83. A log base-2 transformation was performed to generate a more normal distribution. The resultant mean for the transformed number of Sci-Hub download requests in a CBSA 7.83, with a minimum of 2.81 and a maximum of 17.16 (see Table 4).

Table 4. Q2 and Q3 Variable Definitions and Descriptive Statistics, 508 Observations

Variable	Mean	Min	Max	STD	Measurement
Dependent Variable					
$\log_2$(Sci-Hub Requests)	7.23	2.81	17.16	2.75	Log base-2 transformation of Sci-Hub requests within CBSA
Independent Variables					
$\log_2$(Population 25+)	17.23	15.17	23.73	1.59	Log base-2 transformation of total population age 25 and older within CBSA
$\log_2$(PctAdv Degree)	3.15	1.67	4.91	0.58	Log base-2 transformation of percentage of population age 25 and older with an advanced degree within CBSA
$\log_2$(Population 25+ - Faculty)	13.68	10.48	21.01	1.94	Log base-2 transformation of total population age 25 and older minus total faculty within CBSA
$\log_2$(PctAdv Degree-Faculty)	3.05	1.38	4.65	0.58	Log base-2 transformation of percentage of population age 25 and older with an advanced degree minus total faculty within CBSA
$\log_2$(Graduate Students)	11.16	0	17.84	2.69	Log base-2 transformation of total number of graduate students within CBSA
$\log_2$(HEI)	2.45	0	8.18	1.63	Log base-2 transformation of total weighted higher education institutions in CBSA; weights: Associate's *1, Bachelor's *2, Master's*3, Doctorate*4

Independent Variables

Sci-Hub download requests are made by individuals. Those most likely to use the academic literature contained within Sci-Hub are members of the higher education community. Secondly, the technical nature of academic literature suggests the need for advanced subject knowledge to understand the content. To better understand who these people are, this study examined the populations based on the size of the population, the proportion of the population with an advanced degree, how many faculty and graduate students are in the area, and the number and type of higher education institutions in the area.

Using the ACS from the U.S. Census Bureau, the population over 25 in a CBSA was used as the base population who could make Sci-Hub download requests. While the minimum requirement to qualify as a CBSA is a population center of 10,000 people and 228 CBSAs have fewer than 100,000 residents over 25 years of age, some CBSAs like those representing large cities like New York City and Los Angeles are vastly bigger, with populations of 13,918,552 and 8,999,070 respectively. As a result, the distribution for populations is highly right-skewed. The skewedness of the population distribution required a log base-2 transformation to provide a more normal distribution. The resultant mean of the transformed population over 25 is 17.23, with a minimum of 15.17 and a maximum of 23.73 (see Table 4).

As stated above, the technical nature of academic literature tends to require a deeper level of domain knowledge. This type of domain knowledge is frequently conferred with the acquisition of an advanced degree, such as a Master’s degree or doctorate. The total population within a CBSA with an advanced degree, defined in the ACS as having a graduate or professional degree, was divided by the total population to generate a percentage of the population with an advanced degree. The distribution of these percentages is slightly right-skewed with 315 CBSAs having less than 10% of the population with an advanced degree, but 3 areas having more than 25%. A log base-2 transformation provided a more normal distribution. The resultant mean of the transformed percentage of the population over 25 with an advanced degree is 3.15, with a minimum of 1.67 and a maximum of 4.91 (see Table 4).

Faculty, by definition, are possessors of advanced degrees. The total number of faculty in a CBSA, as determined by the IPEDS Institutional Characteristics File and the Employees by Assigned Position File, were subtracted from the ACS population over 25 with an advanced degree. This new total was divided by the total population over 25 to generate the percentage of the population with an advanced degree outside of academia. Based on assertions from publishers (McNutt, 2016), faculty should have access to the academic literature they need through their institutional licenses, this new figure helps to understand how non-academics are using Sci-Hub. As with the percentage of the population with an advanced degree, the percentage of the population with an advanced degree not including faculty is slightly right-skewed. A log base-2 transformation provided a more normal distribution. The resultant mean of the transformed population over 25 minus faculty is 13.68, with a minimum of 10.48 and a maximum of 21.01 (see Table 4).

To better account for how the presence of HEIs are related to Sci-Hub download requests in a region, a weighted total of HEIs was assigned to each CBSA. HEIs are divided into four categories based on the primary degree conferred: Associate, Baccalaureate, Master, and Doctorate (see Appendix A). Associate degree granting institutions, or community colleges, were assigned a multiplier of one; Baccalaureate degree granting institutions, typically liberal arts and comprehensive colleges, were assigned a multiplier of two; Master degree granting institutions were assigned a multiplier of three; Doctoral degree granting institutions, or R1, R2, and R3 universities, were assigned a multiplier of four. The total number of each type of institution was weighted by their multiplier and totaled to generate an HEI variable.

$$\text{HEI} = (\text{Associate}*1) + (\text{Baccalaureate}*2) + (\text{Master}*3) + (\text{Doctorate}*4)$$

The weighted values reflect the differences in resource expenditures by institution type. While the differences in library resource expenditures does not break down where research universities outspend, on average, community colleges at 4:1 ratio, resource expenditures also vary by FTE enrollment. (Phan, Hardesty, & Hug, 2014). The formula used here was created as an estimate to account for the differences between institution types. The HEIs are highly right-skewed with 24 CBSAs having a score of zero and 292 having a score of 5 or less; 5 CBSAs have a score of more than 100. A log base-2 transformation was required to achieve a more normal distribution. The resultant mean of the transformed HEI variable is 2.45, with a minimum of 0 and a maximum of 8.18 (see Table 4).

As previously stated, graduate students are also likely users of academic literature to write papers and perform research (Abbott, 2016). Some graduate students may already have an advanced degree; however, many graduate students working on their first Master's degree or going directly into a doctoral program will not have an advanced degree. With this variation, a simple subtraction from the total population with an advanced degree would not work. The total number of graduate students in a CBSA, as determined by the IPEDS Institutional Characteristics File and the Fall Enrollment File, was generated using the Stata contract command. The distribution of graduate students is highly right-skewed with 116 CBSAs having zero graduate students and 254 having fewer than 1500; 31 CBSAs have more than 100,000 graduate students. A log base-2 transformation was required to achieve a more normal distribution. The resultant mean of the transformed total number of graduate students is 11.16, with a minimum of 0 and a maximum of 17.84 (see Table 4).

Interaction Effects and Collinearity

The correlation matrices (3.5, 3.6) show the correlation between the dependent variables for Q2 and Q3, respectively. The high significance level of the correlations is likely, in part, due to the large sample size. However, the high levels of correlation required testing to determine multicollinearity. A higher density of HEIs in highly populated areas is expected as research institutions are frequently located in cities. Similarly, CBSAs with higher HEI scores are also likely to have more graduate students, especially when accounting for the institution types with the highest weights, Master's and Doctoral granting institutions, are also the institutions which have nearly all of the graduate students. Stata was used to determine the variance inflation factor (VIF); no VIF value greater than 10 was found, so there is likely no problem with multicollinearity (Acock, 2014).

Table 5. Q2 and Q2.1 Correlation Matrix

	Log$_2$(Sci-Hub Requests)	Log$_2$(Population 25+)	Log$_2$(PctAdv Degree)	Log$_2$(Graduate Students)	Log$_2$(HEI)
Log$_2$(Sci-Hub Requests)	1				
Log$_2$(Population 25+)	0.81***	1			
Log$_2$(PctAdv Degree)	0.58***	0.40***	1		
Log$_2$(Graduate Students)	0.68***	0.65***	0.55***	1	
Log$_2$(HEI)	0.78***	0.83***	0.50***	0.73***	1

Table 6. Q3 and Q3.1 Correlation Matrix

	Log$_2$(Sci-Hub Requests)	Log$_2$(Population 25+)	Log$_2$(PctAdv Degree-Faculty)	Log$_2$(Graduate Students)	Log$_2$(HEI)
Log$_2$(Sci-Hub Requests)	1				
Log$_2$(Population 25+)	0.81***	1			
Log$_2$(PctAdv Degree-Faculty)	0.60***	0.44***	1		
Log$_2$(Graduate Students)	0.68***	0.65***	0.52***	1	
Log$_2$(HEI)	0.78***	0.83***	0.50***	0.73***	1

3.3 DATA ANALYSIS

To address the questions outlined at the beginning of this chapter, with the exception of Q1.1, this study used multiple regression (MR). MR allows for multiple independent variables to be incorporated into an equation as a way of understanding the relationships between the variables. The increased number of variables allows for greater explanation of the dependent variable and reduces the possibility of distorting influences of other independent variables (Lewis-Beck, 1980). MR allows for the control of many factors that affect the dependent variable simultaneously, which can be important when relying on nonexperimental data (Wooldridge, 2013), such as the datasets in this study. Specifically, the MR in this study relied on ordinary least squares (OLS) to estimate the model parameters. In MR, OLS regression coefficients "minimize the sum of squared deviations between the model implied scores […] and the observed scores" (Kelley & Maxwell, 2010, p. 286), meaning that the model provides the best fit line between observed values and can be used to predict where other values would fall.

The value of this method is that, when done properly, the relationship between the dependent variable and any one independent variable can be examined by holding the remaining independent variables constant. For example, the relationship between the number of Sci-Hub download requests in a region and the percentage of the population in the region with advanced degrees could be examined by holding the other predictor variables constant.

Variable selection for Question 1 used an exploratory technique described by Kelley and Maxwell (2010). This method was chosen because there was a primary independent variable, academic journal expenditures, and the subsequent independent variables were added to increase the explanatory value of the model. Academic journal expenditures were examined to explore the belief that Sci-Hub requests in the United States are done primarily for convenience instead of need (Bohannon, 2016b; McNutt, 2016). Convenience here is understood through Zipf's Principle of Least Effort (1949); while interlibrary loan may be available to some Sci-Hub users if they are associated with a higher education institution, the delays caused by having to request articles in this way may violate the Principle of Least Effort. Each predictor variable is related to either the perceived access to academic literature (based on journal expenditures) or the perceived need for access to academic literature (based on the size of population of graduate students and faculty, the amount of research conducted, and the presence of a medical school or affiliated hospital). The additional predictors were added one at a time to the basic model to ascertain the improvement of the model with the additional variables; variables accounting for interaction effects were added as needed.

Similarly, Questions 2 and 3 looked at the relationship between the number of Sci-Hub download requests, population size, and the percentage of the population with advanced degrees. Those in possession of advanced degrees are potential users of Sci-Hub based on having the domain knowledge necessary for understanding academic literature, so determining the relationship between the percentage of advanced degree holders, including and excluding faculty, in a CBSA provides insight to how Sci-Hub is used. However, unlike the method used in Question 1, the additional predictors added in Questions 2.1 and 3.1 were added simultaneously to determine the improvement of the model when these variables were accounted for; for instance, the number of graduate students and the number and type of higher education institutions in a region were used as predictors of Sci-Hub usage.

To address Question 1.1, this study employed a simple proportion estimate. Using the sample size calculator from Australia's National Statistical Service (n.d.), a sample of 1065 Sci-Hub requests from the 419,934 total download requests in Q1 was used to obtain a confidence level of 95%. Each Sci-Hub download request contains a DOI that links to a specific journal article. The journal title was extracted from the DOI by performing a search in CrossRef (Crossref, n.d.). A subsequent search of the Online Public Access Catalog (OPAC) of the associated institution's library was performed; a new binary variable that reflected whether that institution subscribes to that journal was then entered into an Excel file and imported into Stata. The results helped shed light on the percentage of Sci-Hub download requests that were for materials that could be accessed through the institutional license of the nearest research university.

3.4 METHOD OF ANALYSIS

3.4.1 Sci-Hub and Academic Libraries

Q1. $$Y = B_0 + B_1 \mathrm{Log}_2(\mathrm{Journals}) + B_2 \mathrm{Log}_2(\mathrm{Graduate/Faculty}) + B_3 \mathrm{Log}_2(\mathrm{Research}) + B_4 \mathrm{Medical/Hospital} + \varepsilon$$

As noted above, multiple regression analysis, specifically OLS, was conducted to understand the relationship between the number of Sci-Hub download requests and institutional journal expenditures, size of graduate student and faculty populations, research funding, and the presence or absence of a medical school or associated hospital. The interrelatedness of the predictor variables could suggest interaction effects between the predictor variables. Each pair of predictor variables was multiplied together to create a byproduct. Models consisting of the independent variable, Sci-Hub requests, the pair of predictor variables, and the byproduct of the predictor variables were run to ascertain if the byproduct was statistically significant. Significant byproducts were analyzed before creating the final model.

Additional predictor variables were added one at a time to gauge what additional explanatory value, if any, the predictor provides. Independent variables that added little explanatory value to the model were dropped. The first independent variable modeled was academic journal expenditures (Log_2(Journals)), as this is the primary relationship described in the research question. The constant (B_0) affords the ability to calculate predicted values. If, as publishers suggest (Bohannon, 2016b), Sci-Hub requests in the United States are a matter of convenience, not necessity, the relationship (B_1) between journal expenditures and Sci-Hub download requests (Y) should be very weak.

After exploring the relationship between Sci-Hub download requests and academic journal expenditures, the additional independent variables were added to the model one at a time. The second variable added was the number of graduate students and faculty (Log_2(Graduate/Faculty)). A larger population of potential users would suggest higher levels of Sci-Hub usage; although, if journal subscriptions are meeting user needs, the number of users should not have a strong effect (B_2) on the number of Sci-Hub download requests. The amount of research (Log_2(Research)) conducted at an institution, as measured by NIH and NSF grant funding, was added to the model next. More research suggests a need for more access to academic literature, which could mean a greater number of Sci-Hub download requests if researchers' information needs are not met through journal subscriptions. Lastly, the presence or absence of a medical school or affiliated hospital Medical/Hospital was added to the model. Health professionals have unique clinical information needs and the presence or absence of a medical school or affiliated hospital could have additional explanatory value (B_4). The included variables were then examined for interaction effects by creating cross-products and modeling.

In the final model, a significant relationship between journal expenditures and Sci-Hub use could provide insight into the convenience versus necessity debate. A negative relationship would suggest necessity is a determining factor; for two otherwise identical institutions, the institution that spends less on journals would expect to have more associated Sci-Hub download requests. Conversely, a positive, significant relationship could suggest convenience as spending more would appear to not deter people from using Sci-Hub. A positive, significant relationship between the number of graduate students and faculty and Sci-Hub use could suggest necessity as a library trying to meet the information needs of more people with the same resources would be likely to fall short of meeting those needs. Similarly, a positive, significant relationship between the amount of research funding received and Sci-Hub use would suggest that identical levels of journal expenditures would not equally suffice two institutions with different levels of research funding. A positive, significant relationship between the medical school or hospital variable and Sci-Hub use could suggest that universities with these affiliations are unable to meet both the academic and clinical information needs of their patrons. However, many universities are located in urban areas with hospitals, regardless of if those hospitals are associated with the university; a negative relationship could suggest that independent hospitals have greater unmet information needs than their counterparts and may be using Sci-Hub to meet those needs. Lastly, the presence of an interaction effect would suggest that how these variables work in concert is more important than how they operate individually.

After the final OLS model was completed, a further examination was performed using a proportion estimate. By limiting the Sci-Hub requests to those in proximity to IRIs, this study determined the percentage of requested documents that were available through the institutions' licenses or if these were for journals outside the collection.

3.4.2 Sci-Hub and Population

$$\text{Q2.}\quad Y = B_0 + B_1\text{Log}_2(\text{Pop 25+}) + B_2\text{Log}_2(\text{Pct Pop 25+ w/ Adv Degree}) + \varepsilon$$

$$\text{Q2.1}\quad Y = B_0 + B_1\text{Log}_2(\text{Pop 25+}) + B_2\text{Log}_2(\text{Pct Pop 25+ w/ Adv Degree}) + B_3\text{Log}_2(\text{Graduate Students}) + B_4\text{Log}_2(\text{HEI}) + \varepsilon$$

$$\text{Q3.}\quad Y = B_0 + B_1\text{Log}_2(\text{Pop 25+}) + B_2\text{Log}_2(\text{Pct Pop 25+ w/ Adv Degree - Faculty}) + \varepsilon$$

$$\text{Q3.1}\quad Y = B_0 + B_1\text{Log}_2(\text{Pop 25+}) + B_2\text{Log}_2(\text{Pct Pop 25+ w/ Adv Degree - Faculty}) + B_3\text{Log}_2(\text{Graduate Students}) + B_4\text{Log}_2(\text{HEI}) + \varepsilon$$

Question 2 examined the relationship between the percentage of residents in a CBSA with an advanced degree and the number of Sci-Hub requests (Y). The total population (Log_2(Pop 25+)) was included to account for the size of the potential user base; the more people in a CBSA, the more potential Sci-Hub users. The coefficient B_1 demonstrates the magnitude of that relationship. The percentage of residents with an advanced degree (Log_2(Pct Pop 25+ w/ Adv Degree)) was included as the primary relationship that was being explored. A larger, positive coefficient (B_2) would suggest that a more educated populace is related with more Sci-Hub download requests. The constant (B_0) provides the ability to calculate predicted variables.

Question 2.1 examined the same relationship as Q2, but added additional independent variables to explore how higher education changed the relationship. Graduate students use academic literature in their classwork and research, so accounting for how many graduate students are in a CBSA (Log_2(Graduate Students)) provided greater explanatory value.

Additionally, the number and type of higher education institutions in a CBSA (Log_2(HEI)) provide additional explanatory value. Positive coefficients for these additional variables (B_3, B_4) would suggest that Sci-Hub download requests have a relationship with the higher education community.

The method of analysis for Q3/3.1 was identical to Q2/2.1, with the exception of subtracting the number of faculty in a CBSA from the population over 25 prior to determining the percentage. If faculty are accessing the information they need through their institutions, the relationships in Q2/2.1 and Q3/3.1 should mirror each other.

In this part of the study, a positive, significant relationship between population size and Sci-Hub use is expected; with all other predictors accounted for, a larger number of people provides more potential Sci-Hub users. A positive, significant relationship between the percentage of the population with an advanced degree and Sci-Hub use could suggest necessity, especially in the models accounting for graduate students and higher education institutions. This finding would suggest that those outside of higher education may be using Sci-Hub to meet their information needs. A positive, significant relationship for this variable when not accounting for higher education institutions and graduate students that changes to negative, or even ceases to be significant, could suggest that most Sci-Hub users come from higher education and the problems of access are primarily associated with colleges and universities.

3.5 STUDY LIMITATIONS

A general limitation of this study was the nature of MR. MR cannot infer causality when the research design is not experimental (Kelley & Maxwell, 2010); as all of the data sources stemmed from existing data sources, this study could only demonstrate that there were relationships between the number of Sci-Hub download requests and any of the independent variables examined. While, the addition of multiple regressors thought to be correlated with the independent variable could increase the explanatory effect of the model, there was no way to control for all potential influencers (Kelley & Maxwell, 2010).

OLS depends on four statistical assumptions: errors have a normal distribution, homogeneity across all regressor values (homoscedasticity), observations are independent of each other, and the relation between the regressors is linear (Kelley & Maxwell, 2010). Despite the transformations performed on the variables to achieve a more normal distribution, none of the distributions is perfectly normal.

Beyond the basic limitations of the method of analysis, each set of research questions in the study has limitations that are specific to it. In the following sections, the specific limitations of each set of research questions are examined.

3.5.1 Sci-Hub and Academic Libraries

The model in Q1 made assumptions about the Sci-Hub download request, specifically that the person who made that request was associated with the nearest R1 or R2 institution. However, not all research is performed at R1 & R2 institutions; Sci-Hub requests associated with one of these institutions could also have come from someone at a non-R1 or R2 institution or someone not in higher education at all.

Limiting Sci-Hub requests to within 10 miles of isolated institutions removed many of the largest research institutions from the study, including nearly 2/3 of all Sci-Hub requests in the U.S. The rationale for this decision was an attempt at avoiding areas such as Pittsburgh where a

Sci-Hub request could reasonably have been associated with the University of Pittsburgh, Carnegie Mellon University, or Duquesne University. However, the sharp drop in the sample size of Sci-Hub download requests should be noted.

Another limitation is the sources for the variable representing research activity. While NIH and NSF are two of the largest funders of research, they are not the only funders and not all research is externally funded. In addition to NIH and NSF, the federal government also funds research through other departments; state governments also supply some external funding. Also, some research funding comes from private non-profit organizations and corporations. None of these other funding sources were included in this study.

Lastly, the 10-mile radius for Sci-Hub download requests provided another limitation. Students in online programs may not live near their institutions and students and faculty may live outside the 10-mile radius used for this study.

3.5.2 Sci-Hub and Population

Using CBSAs as the geographic region presented a limitation since they do not cover the entirety of the United States. Additionally, 3720 of the Sci-Hub download requests coordinates did not return a zip code during the reverse geocoding procedure. The result of these two issues was a total number of included Sci-Hub download requests of 1,100,159, compared to the initial 1,150,963 in the initial dataset. These 50,000 download requests were not accounted for in the study.

Another limitation is how graduate students are counted. Graduate students with an advanced degree may be double counted. For example, a graduate student with a Master’s degree who then enrolled in a doctoral program would be counted in the graduate student variable and the percentage of the population with an advanced degree. As there was no way to ascertain the difference, the assumption was made that they are separate; however, that is not the case for some students.

4.0 RESULTS

This chapter begins the discussion of the multiple regression models that were conducted to answer the research questions posed in this study. The model creation process, interaction effect, and final model for Question 1 are examined first. Next is Question 1.1 which checks a sample of the Sci-Hub download requests to see if the nearby research institutions subscribe to the journal where the article is published. This is followed by an examination of the findings for Questions 2 and 3 and their sub-questions regarding Sci-Hub use and regional characteristics.

4.1 QUESTION 1

How do the academic journal expenditures of research institutions relate to the number of Sci-Hub article requests within a 10-mile radius of the institution?

4.1.1 Model Building and Variable Reduction

In building towards the final model, a series of multiple regressions were run. First each of the variables – journal expenditures, the size of the graduate student and faculty population, the amount of research funding, and the presence of a medical school or hospital – were run one by one as the sole predictor in a regression model (Appendix C.1, Models A, B, C, D). In each of these models, the predictor variable had a significant relationship with Sci-Hub use.

Next each predictor variable was added to the model one by one in building toward the final model. In the modeling journal expenditures on its own had a significant relationship with Sci-Hub usage (Table 7, Model A; Appendix C.1, Model A). When the next model was run adding the size of the graduate student and faculty population, journal expenditures was no longer significant, but the size of the graduate student and faculty population was significant (Table 7, Model B; Appendix C.1, Model E). Next, research funding was added to the model (Table 7, Model C; Appendix C.1, Model F). In this model, research funding was significant, while journal expenditures and the size of the graduate student and faculty population were not. Lastly, the binary variable that represents the presence or absence of a medical school or hospital at an institution was added to the model (Appendix C.1, Model G). While research funding continued to be significant, the addition of the binary variable only added one tenth of one percent to the variance of Sci-Hub download requests explained by the previous model. Due to the low explanatory value of the medical school or hospital variable, it was dropped from the model. The predictor variables used in the final main effects model are journal expenditures, the size of the graduate student and faculty population, and the amount of research funding received (Table 7, Model C; Appendix C.1, Model F).

4.1.1.1 Final Model with Main Effects

Table 7. Coefficients for Q1 Main Effects Regression Models

	Model A		Model B	Model C
log$_2$(Journals)	0.962	***	0.519	0.123
log$_2$(Grad/Faculty)			0.703*	0.544
log$_2$(Research)				0.242**
Intercept	-12.202	**	-11.357**	-6.091
R^2	0.152		0.184	0.231
N	136		136	135

* p<0.05, ** p<0.01, *** p<0.001
See Appendix C.1 for full results

Before examining the final main effects model, a brief explanation of the models building up to it can help provide a better understanding of the relationships between the predictor variables and Sci-Hub use. The model with only journal expenditures as a predictor (Table 7, Model A; Appendix C.1, Model A) shows a positive, significant relationship between journal expenditures and Sci-Hub use. This model provides an R^2 of .152, meaning that this model accounts for approximately 15% of the variation in Sci-Hub use. Journal expenditures were selected as the first predictor variable to examine as they provide the most direct method available for determining the academic literature available to a campus community. In this model, an institution with the mean log$_2$ journal expenditures (22.31) would have an expected 9.26 log$_2$ Sci-Hub download requests; in raw numbers this means that for an institution with $5,214,473 in yearly journal expenditures, we would expect 612 Sci-Hub download requests over the six-month period that Sci-Hub data is available for. However, for an isolated research institution one standard deviation above the mean for journal expenditures, $10,499,061, there are an expected 1199 Sci-Hub download requests.

The number of graduate students and faculty at an institution was chosen as the second predictor variable as it accounts for the number of likely Sci-Hub users. Adding it to the model with only journal expenditures as a predictor (Table 7, Model B; Appendix C.1, Model E) increases the variance accounted for by approximately 3%, to 18.4%. For this new model, holding yearly journal expenditures constant at their mean, a research university with the mean number of graduate students and faculty, 7367, would expect to have 612 Sci-Hub download requests. An isolated research institution with a graduate student and faculty population of 13,691, one standard deviation above the mean, spending the same amount on journal expenditures would expect to have 945 Sci-Hub download requests.

Research funding from NIH and NSF was added to the previous model to account for the perceived need for access to academic literature to perform the literature reviews vital to research. In this final model using only the main effects (Table 7, Model C; Appendix C.1, Model F), only the $\log_2$ transformation of research funding is significant; while the relationships between journal expenditures and the size of the graduate student and faculty population are positive, they are not significant. This model explains 23.1% of the variance in Sci-Hub download requests, a nearly 5% increase from the previous model. The positive relationship between research funding and the number of Sci-Hub download requests suggests that institutions that conduct more research can expect to have more Sci-Hub download requests emanating from near their campus.

To better illustrate the relationship between research funding and Sci-Hub download requests, it is helpful to compare the expected numbers of Sci-Hub download requests at varying levels of research funding. Holding yearly journal expenditures and the size of the graduate student and faculty populations at their respective means, an institution with the mean research funding level, $9,537,039, would expect to have 612 Sci-Hub download requests. For an institution with research funding one standard deviation above the mean, $81,174,204, there are an expected 1028 Sci-Hub download requests. In this example, while research funding rises approximately 8.5 times, the number of Sci-Hub download requests only doubles. As this model only accounts for approximately 23% of the variance in Sci-Hub usage near isolated research institutions, there are clearly additional factors beyond the predictor variables in this model that are contributing to Sci-Hub use.

4.1.2 Interaction

While the final main effects model provides some insight into the relationship between higher education institutions and nearby Sci-Hub download requests, there is a need to examine the relationships further to test for multicollinearity and interaction effects. VIF tests were conducted on the final model of main effects and no multicollinearity was found.

As academic literature is heavily used by graduate students and faculty, especially as part of the research process, there were possibilities of interactions between any and all of the predictors. To further explore the relationship between the predictor variables and Sci-Hub use, a series of interaction effects were tested (Appendix C.1, Models H, I, J); each interaction effect was found to be significant on its own. Since each interaction effect model is significant, regression models were created that included one interaction effect along with the three predictor variables (Appendix C.1, Models K, L, M). The next step was to test a model that included all three interactions (Appendix C.1, Model N); none of the interaction effects variables were significant in this model. The next step was to explore if there was a model where more than one interaction effect was significant, so a series of pair-wise combinations were run (Appendix C.1, Models O, P, Q); none of these models had significant interaction effects. The interaction model chosen was the interaction between the size of the graduate student and faculty population and research funding; this model accounts for the greatest variance of Sci-Hub download requests of all models with a significant interaction, 27%, and the variance accounted for is higher than the model without an interaction.

4.1.3 Final Model with Interaction

Table 8. Coefficients for Q1 Main Effects Regression Models

	Model A
$\log_2$(Journals)	0.14
$\log_2$(Grad/Faculty)	-3.789*
$\log_2$(Research)	-2.100*
Grad/Faculty X Research	0.184**
Intercept	48.279*
R^2	0.270

* $p<0.05$, ** $p<0.01$, *** $p<0.001$
See Appendix C.1, Model M for full results

The final model selected uses the interaction between the size of the graduate student and faculty population and research funding (Table 8, Model A; Appendix C.1, Model M). While examining the main effects provides some understanding of how institutional characteristics are related to the number of Sci-Hub download requests near that institution, an exploration of the interaction between research funding and the size of the graduate student and faculty population provides more insight. In Figure 6, the blue line represents isolated research institutions at the 10th percentile of graduate student and faculty population size, the orange line represents institutions at the mean population size, and the gray line represents institutions at the 90th percentile of size. Holding institutions at the mean for journal expenditures, isolated research institutions at 10th percentile of research funding are at the far left, while institutions at the 90th percentile are on the far right. The number of Sci-Hub download requests are not very different for institutions with low levels of research funding, regardless of the size of their graduate student and faculty populations. However, for institutions with high levels of research funding, the number of Sci-Hub download requests vary greatly based on the number of graduate students and faculty at that institution.

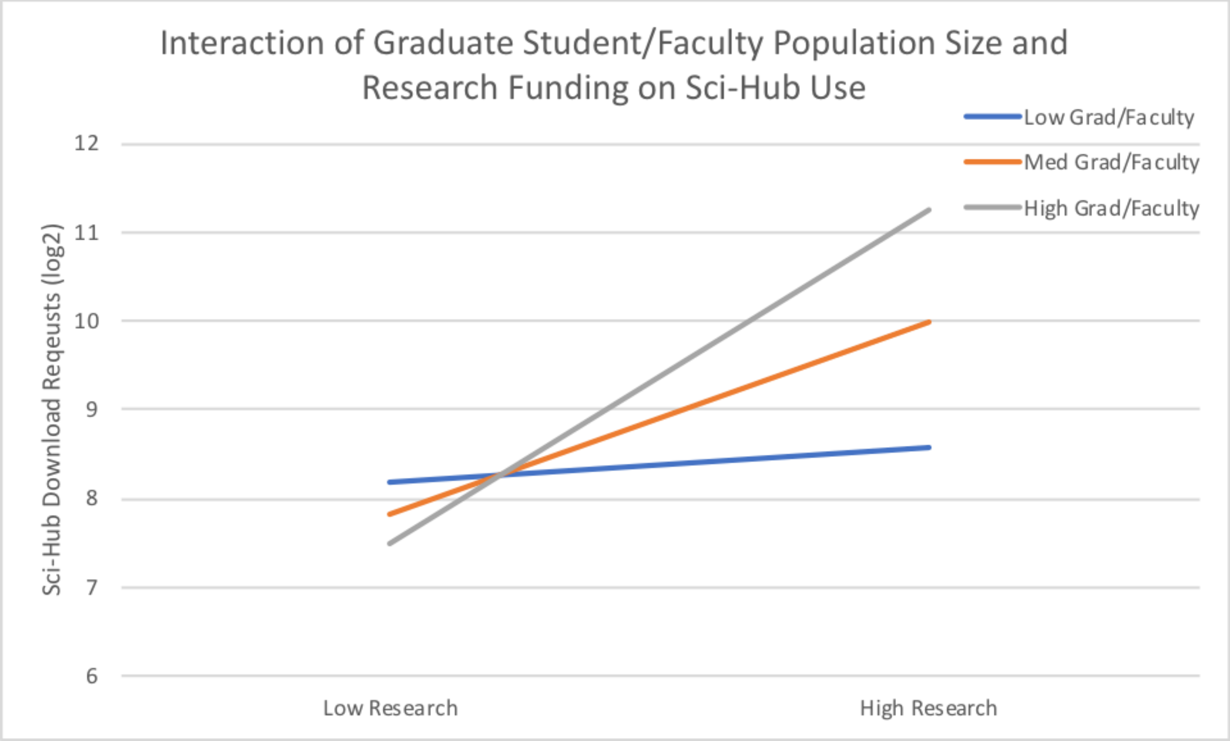

Figure 6. Interaction of Graduate Student/Faculty Population Size and Research Funding on Sci-Hub Use

This figure illustrates the relationship between Sci-Hub use and research funding when comparing universities at the 10th, 50th, and 90th percentiles of graduate student and faculty population size.

The interaction effect is the key predictor of Sci-Hub download requests in this model. The sharp differences in the slopes between isolated research institutions with a small number of graduate students and faculty compared to their counterparts at the high end of that spectrum means that the relationship between Sci-Hub use and higher education institutions differs when these predictor variables are understood together. Table 9 provides a comparison of expected Sci-Hub download requests for institutions at the 10th and 90th percentiles of graduate student and faculty population size and research funding, along with institutions at mean levels for these predictors, while holding journal expenditures at their mean level.

Table 9. Expected Sci-Hub Download Requests by Research Level and Graduate Student/Faculty Population Size

	Low Research	High Research
Low Grad/Faculty	290	379
Med Grad/Faculty	226	1017
High Grad/Faculty	181	2433

For institutions with small populations of graduate students and faculty, the difference in expected Sci-Hub download requests is relatively small even when looking at research funding at the 10th and 90th percentiles where there are less than 100 more Sci-Hub download requests for the high research institutions. However, for institutions with the mean number of graduate students and faculty, there is an expected difference of nearly 800 Sci-Hub download requests between low research and high research institutions. That difference grows even greater when looking at institutions in the 90th percentile of graduate student and faculty population size, where there is an expected difference of over 2200 Sci-Hub download requests compared to their counterparts at the 10th percentile in size.

The University of Minnesota-Twin Cities and the University of Alabama at Birmingham (UAB) provide a good example of this interaction effect. These universities received nearly identical amounts of funding from NIH and NSF, $242 million and $232 million respectively, but the University of Minnesota had 23,389 graduate students and faculty compared to 10,880 at the UAB. There were 2,463 Sci-Hub download requests near the University of Minnesota campus compared to 1,271 near the Birmingham campus. For universities at the lower end of funding from NIH and NSF, Central Michigan University and Baylor University provide a good example. Central Michigan received $1.475 million in research funding, compared to Baylor's $1.124; Central Michigan had 9,767 graduate students and faculty while Baylor had 4,226. Central Michigan had 83 Sci-Hub requests near their campus while Baylor had 93. Central Michigan and UAB also provide a good example of this interaction effect as they have approximately the same number of graduate students and faculty, but vastly different levels of research funding and Sci-Hub downloads near their respective campuses.

Based on this interaction it appears that institutions with low levels of research funding may have lower levels of information needs regardless of the number of graduate students and faculty; without a great need for access to academic literature, there may be less of an impetus to use Sci-Hub. Conversely, with high levels of research funding, there is a greater need for access to academic literature, especially for institutions with a large number of graduate students and faculty. For institutions with small populations of graduate students and faculty, there are fewer potential Sci-Hub users. Additionally, these institutions with smaller populations may have fewer disciplines; a smaller number of disciplines may require a smaller number of journal subscriptions needed to meet users' information needs. However, for institutions with high levels of research funding and large graduate student and faculty populations, there are both more potential Sci-Hub users and more potential disciplines to cover. The journal expenditures for these large, research-intensive universities likely have to cover a wider variety of disciplines and this breadth may come at the expense of depth of coverage. This likely inability to provide the needed depth of coverage for some, or all, disciplines could explain why these institutions are related to higher Sci-Hub use than their peers with less research funding and fewer graduate students and faculty.

4.2 QUESTION 1.1

For Sci-Hub article requests within a 10-mile radius of a research institution, are the requests for articles that appear in journals for which the institution has a subscription?

A random sample of 1065 records from the 419,934 total Sci-Hub download requests within 10 miles of an isolated research institution was created to answer the research question. This sample size provided a 95% margin of error. Each of the DOIs in the sample was searched within CrossRef (Crossref, n.d.) to obtain the journal title for the article. These journal titles were then checked against the library holdings of the isolated research institution associated with that download request. The result is a binary variable for whether that journal was in the library's collection, meaning that members of that campus community would have access to the article in question.

During the course of checking the DOIs of Sci-Hub download requests against library holdings, two problems were encountered. The first issue is that 7 DOIs in the sampled Sci-Hub data did resolve to a record. The second issue is that 4 universities (Louisiana State University, Mississippi State University, Northern Illinois University, and Texas A&M University) do not allow outside users to access their holdings; this affected 14 records. 21 additional records were randomly selected and added to the sample data to get to the total of 1065 records.

Table 10. Percentage of Sci-Hub Requests for Journals with Subscriptions by Nearest IRI

	Frequency	Percent
No Subscription	96	9.01
Subscription	969	90.99

As shown in Table 10, of the sample of Sci-Hub download requests, 969 of the 1065 records were for items in the holdings of the nearest isolated research institution. This tabulation shows that nearly 91% of Sci-Hub download requests near these 136 research institutions were for articles in the library's collection. This figure is predictive for the nearly 420,000 Sci-Hub download requests near isolated research institutions. While this high percentage suggests convenience plays a large role in Sci-Hub use near universities, there is no way to demonstrate conclusively that these download requests were from members of the campus communities. Questions 2 and 3 can help provide insight into Sci-Hub use beyond university campuses.

4.3 QUESTION 2

What is the relationship between the percentage of residents over 25 with an advanced degree in a core-based statistical area (CBSA) and the number of Sci-Hub requests?

Table 11. Q2 Regression Model and Coefficients, N=508

	Model A
$\log_2$(Pop 25+)	1.191***
$\log_2$(Pct Pop 25+ w/ Adv Degree)	1.431***
Intercept	-17.804***
R^2	0.732

* p<0.05, ** p<0.01, *** p<0.001

See Appendix C.2 for full results

Questions 2 and 3, and their sub-questions, expand the geographic areas examined from the 10-mile radii around isolated research institutions to the 508 CBSAs across the United States. Prior to running the regression model, a VIF test was run and found no multicollinearity. Running the regression model to examine the relationship between the number of Sci-Hub download requests within a CBSA and the population size and educational attainment of that population shows positive, significant relationships between both predictor variables (Table 11; Appendix C.2). This means that, when holding the percentage of a CBSA's population with an advanced degree constant, there is an expected increase in Sci-Hub use with larger populations. Similarly, holding population size constant, there is an expected increase in Sci-Hub use with a higher percentage of the population holding advanced degrees. Both of these relationships make sense when examining them; the former says that when there are more people, there tend to be more Sci-Hub download requests, while the latter says that more Sci-Hub use is expected when there are more people with advanced degrees, the target audience for academic literature. Together, these two predictor variables account for over 73% of the variance in Sci-Hub download requests in the United States.

The mean CBSA population is $\log_2(17.23)$ or approximately 153,800 people; for CBSAs with the mean population, there are an expected 150 Sci-Hub download requests when that CBSA also has the mean percent of its population with an advanced degree, 8.87%. However, for a CBSA of the same size, but with a percentage of advanced degree attainment one standard deviation above the mean, 13.31%, there are an expected 268 Sci-Hub download requests. This increase of less than 4.5% in the percentage of the population with an advanced degree raises the number of Sci-Hub download requests from one for every 1025 people to one for every 574 people. The results suggest that people with advanced degrees may have greater needs for access to academic literature than their counterparts without advanced degrees and these information needs may not being met through traditional measures, so they turn to Sci-Hub instead.

4.4 QUESTION 2.1

How does the number and type of higher education institutions, and the total number of graduate students within a CBSA change the above relationship?

Table 12. Q2.1 Regression Model and Coefficients, N=368

	Model A
$\log_2$(Pop 25+)	0.863***
$\log_2$(Pct Pop 25+ w/ Adv Degree)	1.343***
$\log_2$(Graduate Students)	0.106*
$\log_2$(HEI)	0.258
Intercept	-13.573***
R^2	0.722

* $p<0.05$, ** $p<0.01$, *** $p<0.001$
See Appendix C.3 for full results

While the previous model helps to explain the relationship between Sci-Hub download requests and the size and educational attainment of the population, it does not account for the presence of higher education institutions, the primary producers and consumers of academic literature. While this regression model reduces the number of CBSAs from 508 to 368 as some regions do not have higher education institutions or graduate students, it provides needed additional insight on how the presence of higher education changes the relationship between educational attainment and Sci-Hub use. Prior to running the regression model, a VIF test was run and found no multicollinearity.

The relationship between all four predictor variables and Sci-Hub use is positive. However, unlike the other three predictor variables, the variable representing the number and type of higher education institutions is not significant (Table 12; Appendix C.3). For the three significant predictors, the positive relationship makes sense logically and a higher level of any of these predictors is associated with more Sci-Hub download requests in the region. The positive relationships for population size and percentage of population with an advanced degree were explained in the Question 2 section; the positive relationship between the number of graduate students in a CBSA and the number of Sci-Hub download requests makes sense as graduate students are users of academic literature. This model accounts for over 72% of the variance in Sci-Hub download requests in CBSAs with higher education institutions and graduate students.

While holding all other predictors at their respective means, a CBSA with a total number of graduate students one standard deviation above the mean, an increase from 2290 to 14,737 students, would have an expected 197 Sci-Hub download requests compared to the 162 when all predictors are held to their respective means. In this example, an increase of over 12,000 students is only associated with 35 more Sci-Hub download requests. This suggests that while graduate students do play a role in Sci-Hub use, they are not the primary drivers.

Holding population size, the higher education institution variable, and the number of graduate students at their respective means (153,805 people, 5.47, 2290 graduate students), there are an expected 162 Sci-Hub download requests for a region with the mean percentage of the population with an advanced degree. For CBSAs with the percentage of the population with an advanced degree one standard deviation above the mean, from 8.87% to 13.31%, there are an expected 279 Sci-Hub download requests. This 4.4 percentage point increase in the percentage of the population with an advanced degree is associated with more than a 50% increase in Sci-Hub use. This supports the idea that people outside of higher education play an important role in Sci-Hub use.

A regression model of the 140 CBSAs without graduate students (Table 13; Appendix C.6) provides an opportunity to examine the relationship between Sci-Hub use, population size, and advanced degree holders in regions that do not have the large higher education institutions that graduate students are part of. In this model, the population size and the percent of the population with advanced degrees continue to have significant relationships with the number of Sci-Hub download requests. This model accounts for 32% of the variance in Sci-Hub use; the positive relationships reinforces the idea that the presence of a highly educated populace is a driver of Sci-Hub use. The lower variance accounted for suggests other factors in addition to population size and the percentage of the population with an advanced degree are driving Sci-Hub use in these regions. Examining these 140 CBSAs provides a method for confirming the relationships between Sci-Hub use and population size and the percentage of the population with an advanced degree for areas without the academic research associated with graduate students.

Table 13. Regression Model and Coefficients for CBSAs without Graduate Students, N=140

	Model A
$\log_2$(Pop 25+)	0.932***
$\log_2$(Pct Pop 25+ w/ Adv Degree)	0.425*
Intercept	-11.127***
R^2	0.320

* p<0.05, ** p<0.01, *** p<0.001

See Appendix C.6 for full results

4.5 QUESTION 3

What is the relationship between the percentage of residents over 25 (minus the total faculty within the region) with an advanced degree in a CBSA and the number of Sci-Hub requests?

Table 14. Q3 Regression Model and Coefficients, N=508

	Model A
$\log_2$(Pop 25+)	1.186***
$\log_2$(Pct Pop 25+ w/ Adv Degree - Faculty)	1.331***
Intercept	-17.259***
R^2	0.719

* p<0.05, ** p<0.01, *** p<0.001
See Appendix C.4 for full results

The model in Question 3 subtracts the total number of faculty in a CBSA from the number of people with advanced degrees before calculating the percentage of the population with an advanced degree. This change allows for better control of the population with an advanced degree variable to compare with the model in Question 2, providing a mechanism for narrowing advanced degree holders to those not associated with a higher education institution. Prior to running the regression model, a VIF test was run and found no multicollinearity. This model accounts for nearly 72% of the variance in Sci-Hub download requests in CBSAs (Table 14; Appendix C.4).

When holding population size at the mean, 153,800 people aged 25 and over, the expected number of Sci-Hub download requests is 150 for a CBSA with the mean percentage of the population with an advanced degree who are not faculty (3.04%). The mean percentage of residents with an advanced degree is 8.87% when faculty are included and 3.04% when faculty are removed; for a CBSA with the mean population size and percentage with an advanced degree, the expected Sci-Hub download requests is 150 regardless of whether faculty are included or not. For a CBSA with the mean population with a percentage of residents with an advanced degree who are not faculty one standard deviation above the mean, 12.33%, there are an expected 257 Sci-Hub download requests.

The total expected Sci-Hub requests for a CBSA with the percentage of residents with an advanced degree one standard deviation above the mean, 257, is nearly the same as the 268 expected Sci-Hub download requests in a CBSA with advanced educational attainment at one standard deviation above the mean where faculty have not been subtracted. The difference is that in Question 2, the difference between the mean percentage of residents with advanced degrees and one standard deviation above that was less than 4.4%; whereas, in the model for Question 3, the difference is over 9%. This difference suggests that while people who have advanced degrees, but are not faculty, are using Sci-Hub, faculty are still likely contributing to Sci-Hub use.

4.6 QUESTION 3.1

How does the number and type of higher education institutions, and the total number of graduate students within a CBSA change the above relationship?

Table 15. Q3.1 Regression Model and Coefficients, N=368

	Model A
$\log_2$(Pop 25+)	0.781***
$\log_2$(Pct Pop 25+ w/ Adv Degree - Faculty)	1.245***
$\log_2$(Graduate Students)	0.143**
$\log_2$(HEI)	0.292*
Intercept	-12.171***
R^2	0.718

* p<0.05, ** p<0.01, *** p<0.001

See Appendix C.5 for full results

The model for Question 3.1 is likely the most complete for examining the relationship between regional educational attainment and regional Sci-Hub use because it includes graduate students and the higher education variable while keeping faculty members from being double counted as both possessors of advanced degrees and being part of higher education institutions. Prior to running the regression model, a VIF test was run and found no multicollinearity. This model shows a positive, significant relationship for all predictor variables and accounts for nearly 72% of the variance in the Sci-Hub download requests in the 368 CBSAs with higher education institutions and graduate students (Table 15; Appendix C.5).

Unlike the model in Question 2.1, the variable representing the number and type of higher education institutions is significant in this model. HEI scores are the total number of higher education institutions in a CBSA, weighted by institution type. For example, a community college has a value of one, while a university has a value of 4. For CBSAs with the mean levels of the three other predictor variables, a CBSA with a mean HEI score of 2.45 would expect to have 167 Sci-Hub download requests. For CBSAs with an HEI score of 4.08, one standard deviation above the mean, there are an expected 232 Sci-Hub download requests. This suggests that while there are forces beyond higher education that are related to Sci-Hub use, the presence of colleges and universities also factor into Sci-Hub use.

For CBSAs at the mean level of all predictor variables, there are an expected 167 Sci-Hub download requests. For CBSAs where 12.33% of non-faculty residents have advanced degrees, one standard deviation above the mean of 3.04%, there are an expected 277 Sci-Hub download requests. When accounting for faculty, graduate students, and higher education institutions, the percentage of the population with an advanced degree is still positively and significantly related to Sci-Hub use; this suggests a need for academic literature beyond campus communities. For these users, Sci-Hub use may be necessary if they want to retrieve the information they are seeking.

As in Question 2.1, a regression model of the 140 CBSAs without graduate students was run to examine the relationship between Sci-Hub use, population size, and the percentage of residents with an advanced degree in regions that do not have the type of higher education institutions that enroll graduate students. This time faculty have been removed from the residents with advanced degrees. This model shows that both population size and the percentage of the population with an advanced degree are both positively and significantly related to Sci-Hub download requests (Table 16; Appendix C.7). This model accounts for 39% of the variance in Sci-Hub downloads and further supports that there are drivers of Sci-Hub use beyond the higher education community. Again, this additional analysis provides a more complete picture of how Sci-Hub is used in the United States by providing a comparison between regions with the higher levels of research associated with institutions that have graduate students and regions that do not.

Table 16. Regression Model and Coefficients for CBSAs without Graduate Students, N=140

	Model A
$\log_2$(Pop 25+)	1.032***
$\log_2$(Pct Pop 25+ w/ Adv Degree - Faculty)	0.570*
Intercept	-13.045***
R^2	0.390

* p<0.05, ** p<0.01, *** p<0.001
See Appendix C.7 for full results

Lastly, as noted above, not all higher education institutions have graduate students. There are 116 CBSAs without graduate students that have at least one higher education institution. A regression model of these CBSAs that accounts for the presence of higher education institutions (Table 17; Appendix C.8) accounts for over 43% of the variance in Sci-Hub use. The mean HEI score for these 116 CBSAs is 1.66 which is between one community college and either one 4-year college or two community colleges. The positive, significant relationship between the HEI variable and Sci-Hub use shows that even community colleges and 4-year colleges, institutions without graduate students, are related to Sci-Hub use. As with the rest of the models for Questions 2 and 3 and their sub-questions, population size and the percentage of the population with an advanced degree are positive and significant. These relationships suggest that when accounting for population size and smaller, less research-intensive higher education institutions, the percentage of the population with an advanced degree is still associated with Sci-Hub use.

Table 17. Regression Model and Coefficients for CBSAs without Graduate Students, N=116

	Model A
$\log_2$(Pop 25+)	0.952***
$\log_2$(Pct Pop 25+ w/ Adv Degree - Faculty)	0.575**
$\log_2$(HEI)	0.356**
Intercept	-11.954***
R^2	0.433

* p<0.05, ** p<0.01, *** p<0.001
See Appendix C.8 for full results

The regression models in this study demonstrate how Sci-Hub download requests relate to nearby research institutions and their geographic regions more generally. While the sample of Sci-Hub requests that were checked against the holdings of the nearby research institutions show that an overwhelming number of requests could have been met using the institution’s library if the user was a member of that campus community, the regression models suggest a more complex relationship. The following chapter will synthesize the results from both parts of the study and how these results relate to previous research, as well as discuss the implications for research and practice.

5.0 CONCLUSIONS AND IMPLICATIONS

The purpose of this study was to examine Sci-Hub usage in the United States to explore the validity of publisher claims of convenience over necessity. Scholarly journals serve as the primary method for disseminating the findings and theories generated by researchers, especially those at higher education institutions. The current for-profit academic publishing industry limits access to journals based on subscriptions paid by individuals and institutions or through purchasing individual articles. These limitations are imposed by creating paywalls to block unauthorized access to the materials (Estok, 2011). With approximately three-quarters of academic literature residing behind these paywalls (Khabsa & Giles, 2014), potential readers not affiliated with an organization that has subscribed to a desired journal can be stymied as they seek to meet their information needs. Sci-Hub is a repository designed to circumvent paywalls and offer free copies of academic articles, books, and conference proceedings that might otherwise be unavailable. Publishers point to site licenses, open access, and interlibrary loan as alternative mechanisms for accessing academic literature; they contend that, especially in more developed nations, Sci-Hub use is based on convenience, not necessity (McNutt, 2016). However, these claims have not been tested.

Convenience, for the purpose of this study, is based on Zipf's Principle of Least Effort (1949), which posits that people will naturally choose the path of least resistance or effort that does not simultaneously create long term problems greater than the issue at hand. Convenience, specifically for information seeking amongst the higher education community, is further broken down as finding sources that satisfice information needs, ease of access to the information source, and time spent finding the information (Connaway et al., 2011).

For Sci-Hub use to be based on convenience, the ease of access to the information source and the timeliness of the immediate delivery must be the driving forces. This means a user would have access to the article either through institutional subscription or via their library's interlibrary loan program, but chooses to use Sci-Hub because the institution's authentication system may be perceived to be too cumbersome or the turnaround time for interlibrary loan is deemed too long. Conversely, necessity would require the user to not be affiliated with an organization that subscribes to the journal in question and would otherwise be faced with paying upwards of $40 for each article needed. However, with the pressures to conduct research and publish (Liebowitz, 2015), circumventing inconveniences such as multi-factor authentication and interlibrary loan might also be viewed as a necessity. Additionally, for graduate students, deadlines for papers can eliminate interlibrary loan as an option if the information need arises too late. For the purposes of this study, convenience is defined as any Sci-Hub download request for an article that the user could also obtain immediately through an association with an institution that has a site license; conversely, necessity is defined as any Sci-Hub download request for an article that would be inaccessible without either a delay or paying a fee directly to the publisher.

In this study, six months of Sci-Hub server logs released in 2016 by Sci-Hub's creator (Bohannon & Elbakyan, 2016) were combined with data from the Integrated Postsecondary Education Data System (IPEDS) ("The Integrated Postsecondary Education Data System," n.d.), the National Institutes of Health (NIH, n.d.-a), the National Science Foundation (NSF, n.d.), and the U.S. Census Bureau's American Community Survey (US Census Bureau, n.d.-b) to explore the relationship between Sci-Hub use in the United States and the characteristics of the areas surrounding the download requests. Specifically, this study examined how Sci-Hub usage, as measured by the number and location of article download requests, is related to the area surrounding the download requests. This study was broken down into two parts: 1) how Sci-Hub download requests are related to the institutional characteristics of research-intensive universities and 2) how Sci-Hub download requests are related to the population of their geographic location.

In the first part of this study, research-intensive universities were associated with Sci-Hub download requests within a 10-mile radius of the institution. To prevent any overlaps in coverage, any qualifying university within 20 miles of another qualifying institution was removed from the analysis. Multiple regression models were then run to explore the relationship between the number of Sci-Hub download requests associated with an institution and the institution’s journal expenditures, the size of the faculty and graduate student population, and amount of research funding from NIH and NSF. Additionally, interaction effects were tested; the final model included an interaction between the size of an institution’s faculty and graduate student population and the amount research funding.

In the second part of the study, Sci-Hub download requests were split up by the core-based statistical areas (CBSAs) used by the U.S. Census Bureau. Multiple regression models were run to examine the relationship between Sci-Hub use and the size of the CBSA’s population and the percentage of residents with an advanced degree. This initial model was refined by adding variables to represent the presence of higher education institutions in the CBSA. This process was then repeated after subtracting the number of faculty members in a region from the residents with an advanced degree. This additional analysis provided the ability to compare the models, and to better account for the presence of higher education institutions.

Taken together, these regression models provide the data-driven analysis needed to begin to examine the debate between convenience and necessity.

5.1 SUMMARY AND SYNTHESIS OF RESULTS

Based on the results of this study, higher education plays a major role in driving Sci-Hub use in the United States. This same conclusion was reached in past research (Bohannon, 2016b; Greshake, 2016; Kramer, 2016a). This finding is unsurprising as academic journals are marketed specifically for campus communities. At the outset of this study, the relationship between journal expenditures and Sci-Hub use was expected to be the key explanatory variable. The rationale for this expectation was that institutions with similar numbers of graduate students and faculty and similar levels of research funding would require similar access to academic literature. Therefore, institutions with smaller journal expenditures would expect to see more Sci-Hub use nearby than their otherwise identical peers with greater journal expenditures. However, this study found that only when journal expenditures were the sole predictor variable was the relationship significant. In addition, the relationship between journal expenditures and Sci-Hub use was positively related, not negative as had been expected. In this study, the forces driving Sci-Hub usage proved to be more complex than this direct correlation. Specifically, the amount of research funding at an institution was much more important in explaining Sci-Hub use.

In the main effects model for Question 1, the amount of research funding from NIH and NSF was the only significant predictor. The positive relationship showed that institutions with higher levels of research funding were associated with higher levels of Sci-Hub use than their peers with similar levels of journal expenditures and number of graduate students and faculty.

Adding the interaction between research funding and the number of graduate students and faculty provides an even clearer picture of the relationship between Sci-Hub and research institutions. Specifically, the interaction shows that for universities with a small number of graduate students and faculty, the difference in expected Sci-Hub use is fairly small when comparing institutions with low levels of research funding with to those with high levels of research funding. However, for institutions with a large number of graduate students and faculty, the amount of research funding received makes a huge difference. For institutions with high levels of research funding and a large number of graduate students and faculty, there is an expectation of greater Sci-Hub use than that found at institutions with the same level of research funding, but a small number of graduate students and faculty. This means that academic libraries may be able to come close to meeting the needs of a small number of users conducting high levels of research, but as those populations grow, the library likely cannot keep up with demand due to the growing number of journal titles needed (Broad, 1988) and escalating subscription prices (Larivière et al., 2015).

This interaction effect suggests that necessity may be more of a driver of Sci-Hub use than convenience, especially for institutions with higher levels of research funding and a larger number of graduate students and faculty. Bohannon's (2016b) interactive map of the geographic coordinates of all Sci-Hub download requests shows that many of the requests clustered near higher education institutions. Specifically, Bohannon cites the high number of downloads from Columbus, Ohio and East Lansing, Michigan, smaller cities that are the respective homes of Ohio State University and Michigan State University. His conclusion was that the requests are likely coming from members of these higher education communities and their information needs could be met by their respective institutional libraries. In this current study over 90% of the articles requested via Sci-Hub could have been obtained through the nearby university's license to the journal if the requestor was a member of the campus community which could support Bohannon's conclusion. However, for the isolated research institutions in this study, Ohio State is in the 90th percentile in research funding and the 95th percentile in size of the graduate student and faculty populations and Michigan State is in the 75th percentile in research funding and the 90th percentile in the number of graduate students and faculty. The results of this study suggest that these two institutions are prime examples of the necessity over convenience, as they are both at the high ends of the predictor variables that constitute the interaction effect, which would contradict Bohannon's conclusion that the high number of Sci-Hub download requests near Columbus and East Lansing supports the convenience argument. Additionally, neither this study or Bohannon's analysis can guarantee that these requests came from members of the respective campus communities, a point that Kramer (2016b) makes regarding her findings about the percentage of articles related to Utretcht and how many would have been available through the University of Utretcht's site license.

The relationship between the amount of research funding at an institution and Sci-Hub use, especially when accounting for the interaction between research funding and the number of graduate students and faculty, suggests that institutions cannot adequately supply their users with the number of journal subscriptions needed to meet all of their needs and users may turn to alternative methods such as Sci-Hub. Looking at the wider geographic areas where higher education institutions are located, this study found that for CBSAs with graduate students, both the number of graduate students and the weighted higher education variable were positively and significantly related to Sci-Hub use when holding the size of the population and the percentage of the population with advanced degrees who are not faculty constant. This means that CBSAs that have either more colleges and universities or more graduate students would expect to have more Sci-Hub use than their otherwise equal counterparts. For the CBSAs without graduate students that have at least one higher education institution, the higher education institution predictor variable is positively and significantly related to Sci-Hub use. This means that even the presence of more community colleges or 4-year colleges is associated with greater Sci-Hub use, even though these institutions conduct less research. These positive relationships provide additional evidence that Sci-Hub use is connected with higher education; even at less research-intensive institutions users need literature for which may not have access. While the relationship between the weighted higher education variable and the number of graduate students in a region does not directly suggest necessity, Greshake (2016) found the heaviest use times for Sci-Hub were between 9am and 5pm. Since most paywalls are based on IP addresses, if these download requests are coming from academics, they would likely only be using Sci-Hub to access journals their institutions do not subscribe to since using Sci-Hub on campus would create an unnecessary extra step that runs counter to the Principle of Least Effort. Following the Principle of Least Effort, researchers on campus would go directly to the desired article in the simplest manner possible; visiting the Sci-Hub site would create an additional, avoidable task.

While the relationship between higher education and Sci-Hub use was expected, what may be more important is the relationship found between the percentage of the population with an advanced degree and Sci-Hub. A series of multiple regression models were run that included the percentage of the population with an advanced degree in a CBSA. In each of these models, regardless of whether faculty were included or not, the percentage of advanced degree holders was positively and significantly related to Sci-Hub use when accounting for the other predictor variables. This means that even when accounting for the presence of higher education institutions in myriad ways, there appears to be a demand for academic literature outside of the academy. One possible explanation for this is the presence of professions that require access to the highly-specified, technical information found in academic literature. Kramer’s (2016a) analysis of Sci-Hub use in the Netherlands, found that Den Haag, the home of the Dutch parliament and many international organizations, had more download requests than all but two other cities despite not having a university. The findings of this study and Kramer’s suggest that while members of the higher education community utilize academic literature, they are not the only ones who want access.

The current publishing model alienates the producers of academic literature from the product they have created. From this market-driven perspective, academic literature is a commodity like any other product and its readers are not scholars and scientists, but merely customers. As customers, readers are expected to pay for the product like they would in any other commercial transaction, even though the authors who created these products give it away for free as a means to share their findings with the greater academic and scientific community. This study found that the paywalls that separate potential readers from academic literature have created the conditions where Sci-Hub use may be a necessity for some of the people who wish to stay up to date with trends in their fields. Specifically, the interaction between research funding and the number of graduate students and faculty suggests these barriers may keep members of campus communities from the information they need, while the relationship between advanced degree holders and Sci-Hub suggests a similar problem for people outside of higher education.

Overall, the findings of this study align with the limited research previously available: Sci-Hub use in the United States is driven by a combination of necessity and convenience. While publishers may claim that convenience is the driving force, the results of this study suggest that necessity plays a significant role. Members of the higher education community are still likely the largest user base for Sci-Hub in the United States, but the results of this study suggest they are not the only users. The relationship between the percentage of advanced degree holders in a region and the number of Sci-Hub download requests shows unmet information needs outside of higher education. Additionally, for universities with higher levels of research funding that also have a large number of faculty and graduate students, academic libraries may be unable to meet the information needs of their users. Taken together, these findings show the current academic publishing model is inadequate for meeting the information needs of the people.

5.2 IMPLICATIONS FOR RESEARCH

While the findings of this study provide new insight into the relationship between higher education, people with advanced degrees, and Sci-Hub use in the United States, there is still a great deal more to learn about who the typical Sci-Hub user is and why they choose to use Sci-Hub. Specifically, my findings and the findings of Kramer (2016a) show a need to better understand the relationship between advanced degree holders and Sci-Hub use, especially for those outside of higher education. Additional research is needed to understand the information needs of these people. Kramer’s (2016a) suggestion of examining the presence of research organizations in cities to help clarify Sci-Hub use is one possible method as it may help explain Sci-Hub use from researchers and practitioners who are not affiliated with a college or university. The relationship between research funding and Sci-Hub use found in this study suggests that reading relevant literature is a necessary part of the research process and access to that literature is required to accomplish it. A strong relationship between the number and size of non-academic research institutions and Sci-Hub could help explain Sci-Hub use outside of higher education.

Another possible method to better understand the relationship advanced degree holders and Sci-Hub usage found in this study is to examine the prominent industries in a region and the Sci-Hub download requests for articles in those fields. This study did not examine the differences between disciplines. The fields of physics and mathematics have a long history of hosting articles in the pre-print repository arXiv (arXiv, n.d.) which would reduce the need for a service like Sci-Hub. However, multiple studies show chemistry is one of the most popular topics in Sci-Hub (Cabanac, 2016; Greshake, 2017a; Himmelstein et al., 2017). A study examining CBSAs or a smaller geographic region could be done that would split the Sci-Hub download requests by field of study and then compared to the industries in that area could help better understand the relationship between advanced degree holders and Sci-Hub use. Additionally, using a similar model to this study, the DOIs of the Sci-Hub download requests could be broken down by discipline and compared to the academic programs of nearby research institutions to examine the relationship between Sci-Hub use and graduate-level programs.

The high percentage of articles requested via Sci-Hub that would be available through institutional licenses found in this study suggests convenience; however, the relationship between advanced degree holders and Sci-Hub suggests necessity. As previously stated, there is no way to ensure that those Sci-Hub requests were made by members of the nearby campus communities and could be coming from residents who are not a part of the institution. A study to comparing the percentage of people with advanced degrees who live near universities to those who do not live near universities could help provide insight into this possible contradiction.

This study showed relationships between Sci-Hub use, higher education, and people with advanced degrees; however, it cannot assert why people are using Sci-Hub. Recently, Travis (2016) conducted a survey that was admittedly skewed towards Sci-Hub users, but also included people who did not use Sci-Hub. Over 50% of respondents cited lack of journal access as the primary reason for Sci-Hub use; 17% of respondents chose convenience as the primary motive and 23% used Sci-Hub because they object to publisher profits. A more thorough survey of only Sci-Hub users could provide additional insight into the drivers behind Sci-Hub use and the percentage of Sci-Hub users that are affiliated with higher education institutions. These could be compared with the findings in this study to further clarify the convenience versus necessity debate.

Lastly, since Sci-Hub has been forced to continually change their domain name (Silver, 2017), the plugins developed for it (Marcos, 2017; “The Installation of Sci-hub Plugin,” n.d.) have become less stable. Additionally, users now often have to search for the current URL before accessing Sci-Hub. With these developments, the convenience of Sci-Hub may be lower than when the Sci-Hub dataset used in this study was collected. Repeating this study if and when a newer dataset is released could provide further insight into the role of convenience. Comparing the results of this study with data from a less convenient Sci-Hub could lead to further clarification of the whether Sci-Hub is use in the United States is driven by convenience or necessity.

5.3 IMPLICATIONS FOR PRACTICE

For the higher education community, the findings in this study confirm the existence of a known problem: the current academic publishing industry is not meeting the needs of users. In the short term, there is a need to educate faculty and students on the available methods of obtaining articles from journals for which the library does not have a subscription. Advocating for Sci-Hub comes with potential retribution from publishers (Jaschik, 2016; Masnick, 2016; Peet, 2016). However, there are legal means of accessing articles such as interlibrary loan. While interlibrary loan can be less convenient than Sci-Hub, creating greater awareness of it may help users meet their information needs, especially at institutions with larger numbers of graduate students and faculty and higher levels of research funding where this study shows the greatest amount of Sci-Hub use. Additionally, libraries should continue to hone their collection development policies, advocate for open access, and create consortia to have better negotiating positions (A. Pyati, 2007b).

A portion of library funding comes from the indirect costs that colleges and universities set as a negotiated rate with funding agencies. Indirect costs are designed to offset the administrative support, facilities, and equipment that go into supporting research. Library funding is included in these indirect costs (Ledford, 2014). The results of this study suggest that for institutions that conduct high levels of research, especially those with large graduate student and faculty populations, the library is unable to meet their users' information needs. Increasing the share of indirect costs earmarked for libraries could help shrink this gap. While this would not address the long-term trend of skyrocketing costs in journal expenditures, it could provide some short-term relief as libraries work toward better, long-term solutions.

For the Sci-Hub download requests that are based on convenience, where the delays associated with interlibrary loans are deemed by the user to be too long, there is a need for a faster turnaround time to make the existing system more convenient. A new interlibrary loan system, Tipasa, was recently launched and offers libraries the option of creating delivery streams from the lending library directly to the user without any manual interaction required (OCLC, 2017). While a study by Gardner, McLaughlin, and Asher (2017) found no significant correlation between interlibrary loan demand and nearby Sci-Hub download requests, that could mean that users who could not find their desired article may have searched for a different article that would meet their need instead of turning to interlibrary loan. Their study did find that Sci-Hub downloads are moderately correlated with total interlibrary loan articles borrowed, which suggests these institutions have greater unmet journal subscription needs and that users may be filling those gaps with both interlibrary loan and Sci-Hub. Innovative delivery systems like Tipasa, if widely adopted, have the ability to mitigate some of the rationales for Sci-Hub use; however, each interlibrary loan request made by a library comes with a fee paid to the lending library and library budgets are already stretched thin.

Greater interlibrary loan use may help mitigate some of the hindrances users face when searching for academic literature, but it is not a panacea that will fix the larger issue. The academic publishing industry has developed a model that provides them with large profits (DBW, 2014; Larivière et al., 2015), but leaves libraries in a position where they have to continually perform triage on their collections to best meet the needs of their users and stay within their budgets (Fletcher, 2017; Sample, 2012). The current trend of publishers bundling journals to offer better prices per journal has become a common practice; unfortunately, some of the journals in the bundles may not be what an institution needs to meet their users' needs

(Frazier, 2001; Wellcome Trust, 2003; Wellen, 2004). While the greater number of journals may increase overall access, it does not ensure that the access is for the most appropriate journals for that particular library's patrons. Additionally, none of these measures provides any mechanism for reconnecting scholars to their work. The underlying alienation and reification inherent in the current knowledge production and dissemination process would continue unabated even if these stopgap solutions were more widely implemented.

For people outside of higher education who have a need or desire to read academic literature, this study suggests a need for new and innovative ways to deliver needed content. The relationship between people with advanced degrees and Sci-Hub use found in this study suggests that academic literature readers exist beyond higher education campuses and their information needs are not being met. While the open access policies of some federal funding agencies (NIH, n.d.-b; "NSF public access initiative," n.d.) can provide relief for accessing articles funded by public monies, these funding agencies only cover a portion of the disciplines within the realm of academic literature. The results of this study can also be used to expand the conversation around access to information beyond just the higher education community. Critical theory, as has been discussed, calls for action in addition to analysis. From this critical theory perspective, Sci-Hub can be viewed as a direct action in response to the inequities created through the current paywalled content system. The information that people want exists, but the ability to access it is limited by publishers' paywalls. This creates the conditions necessary for people to try to circumvent the system, and they will likely continue to do so until better solutions become available to access academic literature.

5.4 CONCLUSION

While Sci-Hub does not offer a long-term solution to the problem of limitations on access to academic literature, based on the findings of this study, it does demonstrate the need for a larger conversation on restructuring the academic publishing industry so that it meets the needs of all people. To truly fix this problem, there likely needs to be a paradigm shift in how we think of journals and academic publishing. The academic community may need to be willing to distance itself from the traditional mechanics of journal publishing and focus on the purpose: a means of disseminating new theories and research to further the advancement of science and knowledge (Regazzi, 2015). Journals began as a way to compile and share the findings of academic societies; subscription fees went to cover production costs and overhead for the society (Regazzi, 2015). Today, most of the subscription fees go to the profit margins of large publishing companies (Larivière et al., 2015) and the system can actually work against the purpose of sharing knowledge.

Creating mechanisms for sharing ideas should be the starting point for discussing how to reform access to academic literature, even if that means dismantling the current for-profit publishing industry. This study has shown that necessity, both within higher education and outside of it, is likely driving people to use Sci-Hub. Sci-Hub, however, is simply circumventing the barriers created by publishers, it does not offer an alternative method for publishing and sharing. Open access offers an alternative method for equalizing access to the end user. However, there are limits to open access as currently constituted, including the large article processing fees that many publishers charge to keep their profit-margins intact. The current mechanics of publishing still require costs such as copy editors and a robust technological infrastructure. One idea to meet those needs is for colleges and universities to create a consortium to bear these costs together and bypass the publishing industry altogether. A new system built around those who create academic literature would help to connect the scholar with their work and could create a more collaborative environment. While this study does not directly address how to fix the academic publishing market, it does provide additional evidence that something must be done.

APPENDIX A

IPEDS CARNEGIE CLASSIFICATION CODES

Table 18. IPEDS Carnegie Classification Codes

Code Value	Carnegie Classification
1	Associate's--Public Rural-serving Small
2	Associate's--Public Rural-serving Medium
3	Associate's--Public Rural-serving Large
4	Associate's--Public Suburban-serving Single Campus
5	Associate's--Public Suburban-serving Multicampus
6	Associate's--Public Urban-serving Single Campus
7	Associate's--Public Urban-serving Multicampus
8	Associate's--Public Special Use
9	Associate's--Private Not-for-profit
10	Associate's--Private For-profit
11	Associate's--Public 2-year colleges under 4-year universities
12	Associate's--Public 4-year Primarily Associate's
13	Associate's--Private Not-for-profit 4-year Primarily Associate's
14	Associate's--Private For-profit 4-year Primarily Associate's
15	Research Universities (very high research activity)
16	Research Universities (high research activity)
17	Doctoral/Research Universities
18	Master's Colleges and Universities (larger programs)
19	Master's Colleges and Universities (medium programs)
20	Master's Colleges and Universities (smaller programs)
21	Baccalaureate Colleges--Arts & Sciences
22	Baccalaureate Colleges--Diverse Fields
23	Baccalaureate/Associate's Colleges
24	Theological seminaries, Bible colleges, and other faith-related institutions
25	Medical schools and medical centers
26	Other health professions schools
27	Schools of engineering

Table 18 (continued)

28	Other technology-related schools
29	Schools of business and management
30	Schools of art, music, and design
31	Schools of law
32	Other special-focus institutions
33	Tribal Colleges
-3	Not applicable, not in Carnegie universe (not accredited or nondegree-granting)

APPENDIX B

ISOLATED RESEARCH INSTITUTIONS

Arizona State University-Tempe
Auburn University
Augusta University
Ball State University
Baylor University
Brigham Young University-Provo
Brown University
Central Michigan University
Claremont Graduate University
Clemson University
College of William and Mary
Colorado State University-Fort Collins
Cornell University
Dartmouth College
East Carolina University
Florida Atlantic University
Florida Institute of Technology
Illinois State University
Indiana University-Bloomington
Indiana University-Purdue University-Indianapolis
Iowa State University
Jackson State University
Kansas State University
Lehigh University
Louisiana State University and Agricultural & Mechanical College
Miami University-Oxford
Michigan State University
Michigan Technological University

Mississippi State University
Missouri University of Science and Technology
Montana State University
Naval Postgraduate School
New Mexico State University-Main Campus
North Carolina State University at Raleigh
North Dakota State University-Main Campus
Northern Arizona University
Northern Illinois University
Nova Southeastern University
Ohio State University-Main Campus
Ohio University-Main Campus
Oklahoma State University-Main Campus
Old Dominion University
Oregon State University
Pennsylvania State University-Main Campus
Portland State University
Purdue University-Main Campus
South Dakota State University
Southern Illinois University-Carbondale
Stanford University
Stony Brook University
SUNY at Binghamton
Syracuse University
Texas A & M University-College Station
Texas A & M University-Commerce
Texas State University
Texas Tech University
The University of Alabama
The University of Montana
The University of Tennessee-Knoxville
The University of Texas at Austin
The University of Texas at El Paso
The University of Texas at San Antonio
University at Buffalo
University of Alabama at Birmingham
University of Alabama in Huntsville
University of Alaska Fairbanks
University of Arizona
University of Arkansas
University of California-Berkeley
University of California-Davis
University of California-Irvine

University of California-Merced
University of California-Riverside
University of California-Santa Barbara
University of California-Santa Cruz
University of Central Florida
University of Cincinnati-Main Campus
University of Connecticut
University of Dayton
University of Delaware
University of Florida
University of Georgia
University of Hawaii at Manoa
University of Illinois at Urbana-Champaign
University of Iowa
University of Kansas
University of Kentucky
University of Louisiana at Lafayette
University of Louisville
University of Maine
University of Massachusetts-Amherst
University of Massachusetts-Dartmouth
University of Memphis
University of Michigan-Ann Arbor
University of Minnesota-Twin Cities
University of Mississippi
University of Missouri-Columbia
University of Missouri-Kansas City
University of Nebraska-Lincoln
University of Nevada-Las Vegas
University of Nevada-Reno
University of New Hampshire-Main Campus
University of New Mexico-Main Campus
University of North Carolina at Charlotte
University of North Dakota
University of North Texas
University of Northern Colorado
University of Notre Dame
University of Oklahoma-Norman Campus
University of Oregon
University of Puerto Rico-Rio Piedras
University of Rhode Island
University of Rochester
University of South Alabama

University of South Carolina-Columbia
University of South Dakota
University of South Florida-Main Campus
University of Southern Mississippi
University of Tulsa
University of Utah
University of Vermont
University of Virginia-Main Campus
University of Washington-Seattle Campus
University of Wisconsin-Madison
University of Wyoming
Utah State University
Vanderbilt University
Virginia Commonwealth University
Virginia Polytechnic Institute and State University
Wake Forest University
Wayne State University
West Virginia University
Western Michigan University
Wichita State University
Worcester Polytechnic Institute
Yale University

APPENDIX C

C.1 REGRESSION MODELS

Table 19. All Regression Models for Question 1

	Model A			Model B			Model C		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Log2(Journals)	0.962***	(0.196)	0.390						
Log2(Grad/Faculty)				1.120***	(0.220)	0.403			
Log2(Research)							0.360***	(0.063)	0.445
Med School/Hospital									
Journal X Grad/Faculty									
Journal X Research									
Grad/Faculty X Research									
Intercept	-12.202**	(4.376)		-5.132	(2.833)		0.922	(1.467)	
N		136			136			135	
F		24.092			25.918			32.889	
R-squared		0.152			0.162			0.198	

* p<0.05, ** p<0.01, *** p<0.001

Table 19 (continued)

	Model D			Model E			Model F		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Log2(Journals)				0.519	(0.274)	0.211	0.123	(0.304)	0.049
Log2(Grad/Faculty)				0.703*	(0.310)	0.253	0.544	(0.308)	0.195
Log2(Research)							0.242**	(0.085)	0.300
Med School/Hospital	1.465***	(0.409)	0.295						
Journal X Grad/Faculty									
Journal X Research									
Grad/Faculty X Research									
Intercept	8.502***	(0.294)		-11.357**	(4.326)		-6.091	(4.665)	
N		136			136			135	
F		12.820			14.995			13.142	
R-squared		0.087			0.184			0.231	

* p<0.05, ** p<0.01, *** p<0.001

Table 19 (continued)

	Model G			Model H			Model I		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Log2(Journals)	0.125	(0.305)	0.050	-5.920*	(2.448)	-2.403	-2.995*	(1.357)	-1.208
Log2(Grad/Faculty)	0.577	(0.325)	0.207	-10.660*	(4.305)	-3.832			
Log2(Research)	0.252**	(0.091)	0.313				-3.005*	(1.286)	-3.719
Med School/Hospital	-0.166	(0.510)	-0.033						
Journal X Grad/Faculty				0.511**	(0.193)	6.224			
Journal X Research							0.148*	(0.058)	5.086
Grad/Faculty X Research									
Intercept	-6.709	(5.053)		131.545*	(54.164)		68.657*	(29.652)	
N		135			136			135	
F		9.815			12.783			14.580	
R-squared		0.232			0.225			0.250	

* p<0.05, ** p<0.01, *** p<0.001

Table 19 (continued)

	Model J			Model K			Model L		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Log2(Journals)				-5.107*	(2.436)	-2.060	-2.886*	(1.355)	-1.164
Log2(Grad/Faculty)	-3.695*	(1.667)	-1.324	-8.766*	(4.315)	-3.141	0.418	(0.308)	0.150
Log2(Research)	-2.073*	(0.895)	-2.565	0.206*	(0.086)	0.255	-2.709*	(1.300)	-3.354
Med School/Hospital									
Journal X Grad/Faculty				0.420*	(0.194)	5.087			
Journal X Research							0.134*	(0.059)	4.594
Grad/Faculty X Research	0.184*	(0.070)	3.981						
Intercept	49.789*	(20.979)		110.493*	(54.097)		61.450*	(30.031)	
N		135			135		135		
F		16.029			11.303		11.466		
R-squared		0.269			0.258		0.261		

* p<0.05, ** p<0.01, *** p<0.001

Table 19 (continued)

	Model M			Model N			Model O		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Log2(Journals)	0.140	(0.298)	0.057	-2.634	(3.537)	-1.063	-5.165*	(2.430)	-2.083
Log2(Grad/Faculty)	-3.789*	(1.684)	-1.358	-5.275	(5.045)	-1.890	-5.270	(5.045)	-1.889
Log2(Research)	-2.100*	(0.900)	-2.600	-2.178	(1.568)	-2.697	-1.811	(1.523)	-2.242
Med School/Hospital									
Journal X Grad/Faculty				0.142	(0.257)	1.717	0.258	(0.228)	3.130
Journal X Research				0.045	(0.085)	1.531	0.092	(0.070)	3.162
Grad/Faculty X Research	0.184**	(0.071)	3.997	0.112	(0.114)	2.431			
Intercept	48.279*	(21.284)		88.835	(58.915)		112.116*	(53.953)	
N		135			135			135	
F		12.006			8.032			9.447	
R-squared		0.270			0.274			0.268	

* p<0.05, ** p<0.01, *** p<0.001

Table 19 (continued)

	Model P			Model Q		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Log2(Journals)	-1.841	(3.193)	-0.743	-0.998	(1.910)	-0.402
Log2(Grad/Faculty)	-6.418	(4.544)	-2.300	-2.816	(2.335)	-1.009
Log2(Research)	-1.632	(1.174)	-2.020	-2.661*	(1.295)	-3.294
Med School/Hospital						
Journal X Grad/Faculty	0.159	(0.255)	1.924			
Journal X Research				0.051	(0.084)	1.732
Grad/Faculty X Research	0.146	(0.093)	3.175	0.141	(0.101)	3.056
Intercept	81.181	(56.942)		60.935*	(29.923)	
N		135			135	
F		9.637			9.630	
R-squared		0.272			0.272	

* p<0.05, ** p<0.01, *** p<0.001

Table 20. Regression Model for Question 2 (N=508)

	Model A		
Variable	*B*	*SE B*	*β*
Log2(Population 25+)	1.191***	(0.044)	0.688
Log2(Pct Pop 25+ w/ Adv Degree)	1.431***	(0.118)	0.304
Intercept	-17.804***	(0.694)	
R-squared		0.732	
F		690.398	

* p<0.05, ** p<0.01, *** p<0.001

Table 21. Regression Model for Question 2.1 (N=368)

	Model A		
Variable	*B*	*SE B*	*β*
Log2(Population 25+)	*0.863****	*(0.103)*	*0.519*
Log2(Pct Pop 25+ w/ Adv Degree)	1.343***	(0.167)	0.268
Log2(HEI)	0.258	(0.145)	0.124
Log2(Graduate Students)	0.106*	(0.046)	0.102
Intercept	-13.573***	(1.574)	
R-squared		0.722	
F		236.275	

* p<0.05, ** p<0.01, *** p<0.001

Table 22. Regression Model for Question 3 (N=508)

	Model A		
Variable	*B*	*SE B*	*β*
Log2(Population 25+)	1.186***	(0.045)	0.685
Log2(Pct Pop 25+ w/ Adv Degree - Faculty)	1.331***	(0.123)	0.283
Intercept	-17.259***	(0.708)	
R-squared		0.719	
F		646.517	

* p<0.05, ** p<0.01, *** p<0.001

Table 23. Regression Model for Question 3.1 (N=368)

	Model A		
Variable	*B*	*SE B*	*β*
Log2(Population 25+)	0.781***	(0.103)	0.470
Log2(Pct Pop 25+ w/ Adv Degree - Faculty)	1.245***	(0.164)	0.251
Log2(HEI)	0.292*	(0.146)	0.140
Log2(Graduate Students)	0.143**	(0.045)	0.138
Intercept	-12.172***	(1.544)	
R-squared		0.718	
F		230.737	

* p<0.05, ** p<0.01, *** p<0.001

Table 24. Regression Model for CBSAs without Graduate Students (N=140)

	Model A		
	B	*SE B*	*β*
Log2(Population 25+)	0.932***	(0.129)	0.514
Log2(Pct Pop 25+ w/ Adv Degree)	0.425*	(0.164)	0.184
Intercept	-11.127***	(2.090)	
R-squared		0.320	
F		32.167	

* p<0.05, ** p<0.01, *** p<0.001

Table 25. Regression Model for CBSAs without Graduate Students (N=140)

	Model A		
	B	*SE B*	*β*
Log2(Population 25+)	1.032***	(0.140)	0.547
Log2(Pct Pop 25+ w/Adv Degree - Faculty)	0.570*	(0.177)	0.239
Intercept	-13.045***	(2.258)	
R-squared		0.390	
F		36.187	

* p<0.05, ** p<0.01, *** p<0.001

Table 26. Regression Model for CBSAs without Graduate Students, including HEI (N=116)

	Model A		
	B	*SE B*	*β*
Log2(Population 25+)	0.952***	(0.138)	0.505
Log2(Pct Pop 25+ w/ Adv Degree - Faculty)	0.575**	(0.177)	0.241
Log2(HEI)	0.356**	(0.123)	0.211
Intercept	-11.954***	(2.219)	
R-squared		0.433	
F		28.523	

* p<0.05, ** p<0.01, *** p<0.001

BIBLIOGRAPHY

Abbott, A. (2016). The Demography of Scholarly Reading. *The American Sociologist*, *47*(2), 302–318. http://doi.org/10.1007/s12108-016-9315-z

Academia.edu. (n.d.). Retrieved December 12, 2017, from https://www.academia.edu/

Acock, A. C. (2014). *A gentle introduction to Stata* (Fourth). College Station, Texas: A Stata Press Publication, StataCorp LP.

ALA. (n.d.). Interlibrary Loans. Retrieved November 19, 2017, from http://www.ala.org/Template.cfm?Section=interlibraryloan&template=/ContentManagement/ContentDisplay.cfm&ContentID=104201

Anderson, K. (2016). Guest Post: Kent Anderson UPDATED — 96 Things Publishers Do (2016 Edition). Retrieved November 12, 2017, from https://scholarlykitchen.sspnet.org/2016/02/01/guest-post-kent-anderson-updated-96-things-publishers-do-2016-edition/

arXiv. (n.d.). arXiv primer. Retrieved from https://arxiv.org/help/primer

Association of American Publishers. (2017). The Association of American Publishers Welcomes Major Judgment Against "Sci-Hub" Pirate Site. Retrieved November 12, 2017, from http://newsroom.publishers.org/the-association-of-american-publishers-welcomes-major-judgement-against-sci-hub-pirate-site

Association of Research Libraries. (n.d.). History of ARL. Retrieved from http://www.arl.org/about/history#.WAt53ZMrJ24

Association of Research Libraries. (2002). *ARL statistics: Published annually since 1962*. Washington, D.C. Retrieved from http://www.arl.org/storage/documents/publications/arl-statistics-2001-02.pdf

Basic Classification. (n.d.). Retrieved September 17, 2017, from http://carnegieclassifications.iu.edu/classification_descriptions/basic.php

Basic Classification Methodology. (n.d.). Retrieved October 1, 2017, from http://carnegieclassifications.iu.edu/methodology/basic.php

Birkinshaw, J. (2014, June). Beyond the information age. *Wired*. Retrieved from https://www.wired.com/insights/2014/06/beyond-information-age/

Björk, B. (2016). The open access movement at a crossroad: Are the big publishers and academic social media taking over? . *Learned Publishing* . Chichester, UK : John Wiley & Sons, Ltd . http://doi.org/10.1002/leap.1021

Bohannon, J. (2016a). The frustrated science student behind Sci-Hub. *Science*. http://doi.org/10.1126/science.aaf5675

Bohannon, J. (2016b). Who’s downloading pirated papers? Everyone. *Science*. http://doi.org/10.1126/science.aaf5664

Bohannon, J., & Elbakyan, A. (2016, April 28). Data from: Who’s downloading pirated papers? Everyone. *Science*. Dryad Data Repository. http://doi.org/doi:10.5061/dryad.q447c

Bombardieri, M. (2014, March 30). Aaron Swartz and MIT: The inside story. *The Boston Globe*. Boston. Retrieved from https://www.bostonglobe.com/metro/2014/03/29/the-inside-story-mit-and-aaron-swartz/YvJZ5P6VHaPJusReuaN7SI/story.html

Broad, W. J. (1988). Science can’t keep up with flood of new journals: at least 40,000 journals are now published yearly, producing a million articles. *New York Times*. New York, N.Y: New York Times Company.

Bronner, S. E. (2011). *Critical theory : a very short introduction*. Oxford University Press.

Cabanac, G. (2016). Bibliogifts in LibGen? A study of a text-sharing platform driven by biblioleaks and crowdsourcing. *Journal of the Association for Information Science and Technology*, *67*(4), 874–884. http://doi.org/10.1002/asi.23445

Case, D. O., & Given, L. M. (2016). *Looking for information: a survey of research on information seeking, needs, and behavior* (Fourth). Bingley, UK: Emerald.

Cavanaugh, A. (n.d.). Henry Buhl Library: ILL &amp; Doc Delivery: Interlibrary Loan/Document Delivery Home. Retrieved from http://hbl.gcc.libguides.com/ill

Chang, Y.-W. (2016). Influence of human behavior and the principle of least effort on library and information science research. *Information Processing & Management*, *52*(4), 658–669. http://doi.org/10.1016/J.IPM.2015.12.011

Chawla, D. (2017). Publishers take ResearchGate to court, alleging massive copyright infringement. *Science*. http://doi.org/10.1126/science.aaq1560

Choo, C. W., Detlor, B., & Turnbull, D. (2000). *Web work : information and seeking knowledge work on the World Wide Web* (Vol. 1). Boston;Dordrecht, The Netherlands; Kluwer Academic Publishers.

Cochran, A. (2016). A Funny Thing Happened on the Way to OA. Retrieved November 12, 2017, from https://scholarlykitchen.sspnet.org/2016/02/25/a-funny-thing-happened-on-the-way-to-oa/

Colbert, P. (n.d.). Interlibrary Loan @ Pitt: Home. Retrieved November 19, 2017, from http://pitt.libguides.com/ill

Cold Spring Harbor Laboratory. (n.d.). About bioRxiv. Retrieved from http://biorxiv.org/about-biorxiv

Connaway, L. S., Dickey, T. J., & Radford, M. L. (2011). “If it is too inconvenient I’m not going after it:” Convenience as a critical factor in information-seeking behaviors. *Library & Information Science Research*, *33*(3), 179–190. http://doi.org/10.1016/j.lisr.2010.12.002

Crossref. (n.d.). Crossref. Retrieved April 1, 2018, from https://www.crossref.org/

D-Scholarship@Pitt. (n.d.). About the repository. Retrieved from https://d-scholarship.pitt.edu/information.html

DBW. (2014). How much money the biggest publishers actually make. Retrieved from http://www.digitalbookworld.com/2014/how-much-money-the-biggest-publishers-actually-make/

DeAngelo, L. (2010). Preparing for the PhD at a comprehensive institution: Perceptions of the “barriers.”. *Journal of the Professoriate, 3*(2).

Dorsch, J. L., & Pifalo, V. (1997). Information needs of rural health professionals: a retrospective use study. *Bulletin of the Medical Library Association*, *85*(4), 341.

Dougherty, R. M. (1989). Are libraries hostage to rising serials costs? *The Bottom Line*, *2*(4), 25–27. http://doi.org/10.1108/eb025197

Easton, C. (1999). The sure things in life serials crisis and cancellation information on the world wide web. *Serials Review*, *25*(2), 69–75. http://doi.org/10.1016/S0098-7913(99)00008-8

Elbakyan, A. (2015). [Untitled]. Retrieved from https://torrentfreak.com/images/sci-hub-reply.pdf

Elmes, J. (2017). Elsevier victory over Sci-Hub “shows research is corporate asset.” Retrieved November 5, 2017, from https://www.timeshighereducation.com/news/elsevier-victory-over-sci-hub-shows-research-corporate-asset

Elsevier. (n.d.). Federated authentication through SAML. Retrieved November 26, 2017, from https://www.elsevier.com/solutions/sciencedirect/support/federated-authentication-through-saml

Elsevier. (2016). Elsevier OA price list. Retrieved from https://www.elsevier.com/__data/promis_misc/j.custom97.pdf

Esposito, J. (2016). Sci-Hub and the Four Horsemen of the Internet. Retrieved December 11, 2017, from https://scholarlykitchen.sspnet.org/2016/03/02/sci-hub-and-the-four-horsemen-of-the-internet/

Estok, D. (2011). Paywalls. *Journal of Professional Communication*, *1*(1).

F1000Research. (n.d.). FAQs. Retrieved from https://f1000research.com/faqs

Fisher, J. H. (2008). Scholarly publishing re-invented: Real costs and real freedoms. *Journal of Electronic Publishing*, *11*(2).

Fletcher, R. (2017). U of C axes hundreds of journal subscriptions as "big 5" publishers jack up prices. Retrieved March 19, 2017, from http://www.cbc.ca/news/canada/calgary/university-calgary-cancels-journal-subscriptions-2017-1.3942774

Frazier, K. (2001). The librarians' dilemma: Contemplating the costs of the "big deal." *D-Lib Magazine*, *7*(3). Retrieved from http://www.dlib.org/dlib/march01/frazier/03frazier.html

Fromm, E., & Marx, K. (1966). *Marx's concept of man* (Vol. M116). New York: F. Ungar.

Gardner, C. C., & Gardner, G. J. (2016). Fast and Furious (at Publishers): The Motivations behind Crowdsourced Research Sharing. *College & Research Libraries*, crl16-840.

Gardner, C., & Gardner, G. (2015). Bypassing interlibrary loan via Twitter: an exploration of #icanhazpdf requests. In *ACRL 2015*. Portland, OR. Retrieved from http://eprints.rclis.org/24847/2/gardner.pdf

Gardner, G. J., McLaughlin, S. R., & Asher, A. D. (2017). Shadow libraries and you: Sci-hub usage and the future of ill.

Gennaro, R. De. (1977). Escalating journal prices: Time to fight back. *American Libraries*. American Library Association.

Google Scholar Support for Libraries. (n.d.). Retrieved October 29, 2017, from https://scholar.google.com/intl/en/scholar/libraries.html

Greco, A. N. (2015). Academic libraries and the economics of scholarly publishing in the twenty-first century: portfolio theory, product differentiation, economic rent, perfect price discrimination, and the cost of prestige. *Journal of Scholarly Publishing*, *47*(1), 1–43. Retrieved from http://10.0.12.66/jsp.47.1.01

Greshake, B. (2016, May 30). Correlating the Sci-Hub data with World Bank Indicators and Identifying Academic Use. http://doi.org/10.15200/winn.146485.57797

Greshake, B. (2017a). Looking into Pandora's Box: The Content of Sci-Hub and its Usage. *F1000Research*, *6*, 541. http://doi.org/10.12688/f1000research.11366.1

Greshake, B. (2017b, April 5). Data And Scripts For Looking Into Pandora'S Box: The Content Of Sci-Hub And Its Usage. http://doi.org/10.5281/zenodo.472493

Gunnarsdóttir, K. (2005). Scientific journal publications: on the role of electronic preprint exchange in the distribution of scientific literature. *Social Studies of Science*, *35*(4), 549–579. http://doi.org/10.1177/0306312705052358

Hahnel, M. (2017). List of DOIs of papers collected by SciHub. *Figshare*. Figshare. http://doi.org/10.6084/m9.figshare.4765477.v1

Harvey, C. (2016, May 18). Climate change doubters really aren't going to like this study.

Washington Post. Retrieved from https://www.washingtonpost.com/news/energy-environment/wp/2016/05/18/climate-change-doubters-really-really-arent-going-to-like-this-study/?utm_term=.74820ffa2687

Heathers, J. (2016). Why Sci-Hub Will Win. Retrieved November 12, 2017, from https://medium.com/@jamesheathers/why-sci-hub-will-win-595b53aae9fa

Himmelstein, D. S., Romero, A. R., McLaughlin, S. R., Tzovaras, B. G., & Greene, C. S. (2017). Sci-Hub provides access to nearly all scholarly literature. *PeerJ Preprints*. http://doi.org/10.7287/peerj.preprints.3100v2

Horkheimer, M. (1982). *Critical theory: selected essays*. New York: Continuum Pub. Corp.

Ivins, O. (1989). Serials prices. *Serials Review*, *15*(3), 71,78-95,95. http://doi.org/10.1016/0098-7913(89)90042-7

Jaschik, S. (2016, August). Supporting Sci-Hub vs. explaining Sci-Hub. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2016/08/08/letter-publishers-group-adds-debate-over-sci-hub-and-librarians-who-study-it

Jongejan, A. (2003). Why commercial publishers are good for research. *Research Informatioin*. Retrieved from https://web-beta.archive.org/web/20031202195526/http://www.researchinformation.info/riaut03elsevier.html

Kelley, K., & Maxwell, S. E. (2010). Multiple regression. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 281–297). New York: Routledge.

Khabsa, M., & Giles, C. L. (2014). The Number of Scholarly Documents on the Public Web. *PLoS ONE*, *9*(5), e93949. http://doi.org/10.1371/journal.pone.0093949

Kmietowicz, K., Wise, A., Chan, L., Kirsop, B., Arunachalam, S., Packenham-Walsh, N., … Smith, R. (2011). On the path to global open access: a few more miles to go. *PLoS Medicine*, *8*(3), e1001014. http://doi.org/10.1371/journal.pmed.1001014

Kmietowicz, Z. (2011). Publishers withdraw 2500 journals from free access scheme in Bangladesh. *BMJ*, *342*. Retrieved from http://www.bmj.com/content/342/bmj.d196.abstract

Kocken, G. J., & Wical, S. H. (2013). "I've Never Heard of It Before": Awareness of Open Access at a Small Liberal Arts University. *Behavioral & Social Sciences Librarian*, *32*(3), 140–154. http://doi.org/10.1080/01639269.2013.817876

Kramer, B. (2016a). Sci-Hub: access or convenience? A Utrecht case study (part 1). Retrieved May 11, 2017, from https://im2punt0.wordpress.com/2016/06/20/sci-hub-utrecht-case-study-part-1/

Kramer, B. (2016b). Sci-Hub: access or convenience? A Utrecht case study (part 2). Retrieved May 11, 2017, from https://im2punt0.wordpress.com/2016/06/20/sci-hub-access-or-convenience-a-utrecht-case-study-part-2/

Kwon, D. (2017). Publishers' Legal Action Advances Against Sci-Hub. Retrieved November 12, 2017, from http://www.the-scientist.com/?articles.view/articleNo/50361/title/Publishers--Legal-Action-Advances-Against-Sci-Hub/

Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PloS One*, *10*(6), e0127502. http://doi.org/10.1371/journal.pone.0127502

Ledford, H. (2014). Indirect costs: Keeping the lights on. *Nature*, *515*(7527), 326–329. http://doi.org/10.1038/515326a

Lewandowsky, S., Ballard, T., Oberauer, K., & Benestad, R. (2016). A blind expert test of contrarian claims about climate data. *Global Environmental Change*, *39*, 91–97.

http://doi.org/10.1016/j.gloenvcha.2016.04.013

Lewis-Beck, M. S. (1980). *Applied regression: an introduction* (Vol. ser. no. 0). Newbury Park, Calif: Sage Publications.

Liebowitz, J. (2015). *A guide to publishing for academics: inside the publish or perish phenomenon*. Boca Raton, Florida: CRC Press.

Machin-Mastromatteo, J. D., Uribe-Tirado, A., & Romero-Ortiz, M. E. (2016). Piracy of scientific papers in Latin America: An analysis of Sci-Hub usage data. *Information Development*, *32*(5), 1806–1814. http://doi.org/10.1177/0266666916671080

Marcos, A. (2017). Sci Hub it! Retrieved October 29, 2017, from https://addons.mozilla.org/en-US/firefox/addon/sci-hub-it/

Masnick, M. (2016). Publishers Association Sends Whiny Complaint Letter To Dean After Academic Librarian Discusses Sci-Hub. Retrieved from https://www.techdirt.com/articles/20160809/00440235190/publishers-association-sends-whiny-complaint-letter-to-dean-after-academic-librarian-discusses-sci-hub.shtml

McKenzie, L. (2017). Sci-Hub’s cache of pirated papers is so big, subscription journals are doomed, data analyst suggests. *Science*. http://doi.org/10.1126/science.aan7164

McLaughlin, S. (2017). Alexandra Elbakyan just blocked all of Russia from accessing Sci-Hub. Here’s the message that pops up: original & translated by Google.

McNutt, M. (2016). My love-hate of Sci-Hub. *Science, 352*(6285).

Multifactor Authentication at Pitt. (n.d.). Retrieved November 26, 2017, from http://technology.pitt.edu/services/multifactor-authentication-pitt

Murphy, K. (2016, March 12). Should all research papers be free? *New York Times*. Retrieved from http://www.nytimes.com/2016/03/13/opinion/sunday/should-all-research-papers-be-free.html?_r=0

National Library of Medicine. (n.d.). PMC overview. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/about/intro/

National Science Foundation. (n.d.-a). NSF public access repository. Retrieved from http://par.nsf.gov/

National Science Foundation. (n.d.-b). Public access to results of NSF-funded researh. Retrieved from https://www.nsf.gov/news/special_reports/public_access/

NIH. (n.d.-a). NIH Awards by Location and Organization. Retrieved September 17, 2017, from https://www.report.nih.gov/award/index.cfm

NIH. (n.d.-b). NIH public access policy details. Retrieved from https://publicaccess.nih.gov/policy.htm

Nolan, H. (2012). Just because you don’t like a study doesn’t mean it is wrong. Retrieved January 1, 2016, from http://gawker.com/5925897/just-because-you-dont-like-a-study-doesnt-mean-it-is-wrong

NPR. (2016). Expensive Journals Drive Academics To Break Copyright Law. Retrieved November 12, 2017, from https://www.npr.org/2016/02/20/467468361/expensive-journals-drive-academics-to-break-copyright-law

NSF. (n.d.). Rankings by total R&amp;D expenditures. Retrieved September 17, 2017, from https://ncsesdata.nsf.gov/profiles/site?method=rankingBySource&ds=herd

NSF public access initiative. (n.d.). Retrieved January 31, 2015, from https://www.nsf.gov/about/budget/fy2014/pdf/45_fy2014.pdf

OCLC. (2017). OCLC introduces “Tipasa” interlibrary loan management system. Retrieved April 23, 2018, from https://www.oclc.org/en/news/releases/2017/201701dublin.html

Office for Scholarly Communication. (n.d.). About DASH. Retrieved from https://dash.harvard.edu/docs/about/

Open Access. (n.d.). Retrieved from https://www.elsevier.com/about/open-science/open-access

OpenCage Data. (n.d.). OpenCage Geocoder. Retrieved September 24, 2017, from https://geocoder.opencagedata.com/

Oxenham, S. (2016). Meet the Robin Hood of Science. http://doi.org/10.3389/fnhum.2013.00291

Pascarelli, A. M. (1990). Coping strategies for libraries facing the serials pricing crisis. *Serials Review*, *16*(1), 75–80. http://doi.org/http://dx.doi.org/10.1016/0098-7913(90)90045-Z

PeerJ. (n.d.). PeerJ preprints. Retrieved from https://peerj.com/preprints/

Peet, L. (2016). Sci-Hub controversy triggers publishers' critique of librarian. Retrieved from http://lj.libraryjournal.com/2016/08/copyright/sci-hub-controversy-triggers-publishers-critique-of-librarian/#_

Peters, J. (2016). The lawsuit against Sci-Hub begs the question: Why are academic journals so expensive, anyway? Retrieved November 12, 2017, from http://www.slate.com/articles/health_and_science/science/2016/04/the_lawsuit_against_sci_hub_begs_the_question_why_are_academic_journals.html

Peters, M., Lankshear, C., & Olssen, M. (2003). *Critical theory and the human condition: founders and praxis* (Vol. 168.;168;). New York: P. Lang.

Phan, T., Hardesty, L., & Hug, J. (2014). *Academic libraries: 2012 first look*. Washington, D.C: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2014/2014038.pdf

Picard, R. (n.d.). help geonear. Retrieved September 1, 2017, from http://fmwww.bc.edu/repec/bocode/g/geonear.html

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., … Haustein, S. (2017). The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ Preprints*. http://doi.org/10.7287/peerj.preprints.3119v1

Powell, W. W., & Snellman, K. (2004). The knowledge economy. *Annual Review of Sociology*, *30*(1), 199–220. http://doi.org/10.1146/annurev.soc.29.010202.100037

Principles and guidelines. (2016). Retrieved January 26, 2017, from http://www.springernature.com/gp/researchers/sharedit/principles

Pyati, A. (2007a). A critical theory of open access: Libraries and electronic publishing. *First Monday*, *12*(10).

Pyati, A. (2007b). A critical theory of open access: Libraries and electronic publishing. *First Monday*, *12*(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/1970/1845

Pyati, A. K. (2006). Critical Theory and Information Studies: A Marcusean Infusion. *Policy Futures in Education*, *4*(1), 83–89. http://doi.org/10.2304/pfie.2006.4.1.83

Read the Budapest open access initiative. (2002). Retrieved from http://www.budapestopenaccessinitiative.org/read

Regazzi, J. J. (2015). Scholarly communications: a history from content as king to content as kingmaker. New York;Boulder;Lanham;London; Rowman & Littlefield.

Research4Life home. (n.d.). Retrieved January 22, 2017, from http://www.research4life.org/

ResearchGate. (n.d.). Retrieved December 12, 2017, from https://www.researchgate.net/

RoyalKoala23. (n.d.). About /r/Scholar : Scholar. Retrieved November 19, 2017, from https://www.reddit.com/r/Scholar/comments/716ov3/about_rscholar/

Ruff, C. (2016, May 13). What do the authors of Sci-Hub's most-downloaded articles think about Sci-Hub? http://doi.org/10.1056/NEJMoa1402121

Saleh, A. A., Ratajeski, M. A., & Ladue, J. (2014). Development of a Web-based repository for sharing biomedical terminology from systematic review searches: a case study. *Medical Reference Services Quarterly*, *33*(2), 167–178. http://doi.org/10.1080/02763869.2014.897518

Sample, I. (2012). Harvard University says it can't afford journal publishers' prices. Retrieved November 24, 2017, from https://www.theguardian.com/science/2012/apr/24/harvard-university-journal-publishers-prices

Schiermeier, Q. (2017). Pirate paper website Sci-Hub dealt another blow by US courts. *Nature*. http://doi.org/10.1038/nature.2017.22971

Scholar. (n.d.). Retrieved November 19, 2017, from https://www.reddit.com/r/Scholar/

Schwieder, D. (2016). Low-Effort Information Searching: The Heuristic Information-Seeking Toolkit. *Behavioral & Social Sciences Librarian*, *35*(4), 171–187. http://doi.org/10.1080/01639269.2017.1289019

Sci-Hub. (n.d.). Retrieved March 18, 2017, from http://sci-hub.cc/

Sherman, D. (2016). Adorno's negative dialectics. *Philosophy Compass*. Ithaca: Wiley Subscription Services, Inc. http://doi.org/10.1111/phc3.12328

Shrauger, K., & Scharf, M. (n.d.). Exploring the Value of Interlibrary Loan. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/confsandpreconfs/2017/ExploringValueofInterlibraryLoan.pdf

Silver, A. (2017). Sci-Hub domains inactive following court order. Retrieved April 3, 2018, from https://www.theregister.co.uk/2017/11/23/sci_hubs_become_inactive_following_court_order/

SocArXiv. (n.d.). SocOpen: the SocArXiv blog. Retrieved from https://socopen.org/

SPARC. (n.d.). Open access. Retrieved from http://sparcopen.org/open-access/

SPARC. (n.d.). SPARC: Advancing open access, open data, open education. Retrieved from http://sparcopen.org/

SPARC. (n.d.). Who we are. Retrieved from http://sparcopen.org/who-we-are/

Springer Open Choice. (n.d.). Retrieved from https://www.springer.com/gp/open-access/springer-open-choice

Standish, R. (2017). The World's Largest Free Scientific Resource Is Now Blocked in Russia – Foreign Policy. Retrieved November 20, 2017, from http://foreignpolicy.com/2017/09/06/the-worlds-largest-free-scientific-resource-is-now-blocked-in-russia/

Statistics, c=AU;o=Australian G. B. of. (n.d.). Sample Size Calculator. Retrieved from http://www.nss.gov.au/nss/home.nsf/pages/Sample+size+calculator

Suber, P. (2015). Open access overview. Retrieved from http://legacy.earlham.edu/~peters/fos/overview.htm

Swartz, A. (2008). Full text of &quot;Guerilla Open Access Manifesto&quot; Retrieved November 19, 2017, from https://archive.org/stream/GuerillaOpenAccessManifesto/Goamjuly2008_djvu.txt

Tenopir, C., & King, D. W. (2001). Electronic journals: how user behaviour is changing.

Tenopir, C., Volentine, R., & King, D. W. (2012). Scholarly Reading and the Value of Academic Library Collections: results of a study in six UK universities. *Insights*, *25*(2), 130–149. http://doi.org/10.1629/2048-7754.25.2.130

The Installation of Sci-hub Plugin. (n.d.). Retrieved October 29, 2017, from https://chempeng.github.io/2017/05/21/Sci-hub-extension/

The Integrated Postsecondary Education Data System. (n.d.). Retrieved from http://nces.ed.gov/ipeds/

Timberg, C. (2016, November 24). Russian propaganda effort helped spread “fake news” during election, experts say. *Washington Post*. Retrieved from https://www.washingtonpost.com/business/economy/russian-propaganda-effort-helped-spread-fake-news-during-election-experts-say/2016/11/24/793903b6-8a40-4ca9-b712-716af66098fe_story.html

Travis, J. (2016). In survey, most give thumbs-up to pirated papers. *Science*. http://doi.org/10.1126/science.aaf5704

U.S. Department of Housing and Urban Development. (n.d.). HUD USPS ZIP Code Crosswalk Files. Retrieved September 24, 2017, from https://www.huduser.gov/portal/datasets/usps_crosswalk.html

United States District Court Southern District Of New York. Elsevier Inc. et al v. Sci-Hub et al (2015). Retrieved from https://www.unitedstatescourts.org/federal/nysd/442951/1-0.html

US Census Bureau. (n.d.). 2010 Geographic Terms and Concepts - Core Based Statistical Areas and Related Statistical Areas. Retrieved October 1, 2017, from https://www.census.gov/geo/reference/gtc/gtc_cbsa.html

US Census Bureau. (n.d.). American Community Survey (ACS). Retrieved September 17, 2017, from https://www.census.gov/programs-surveys/acs/

US Census Bureau. (n.d.). US Census Bureau 2010 ZIP Code Tabulation Area (ZCTA) Relationship File Layouts and Contents. Retrieved September 17, 2017, from https://www.census.gov/geo/maps-data/data/zcta_rel_layout.html

Vogel, G., & Kupferschmidt, K. (2017). A bold open-access push in Germany could change the future of academic publishing. *Science*. http://doi.org/10.1126/science.aap7562

Wellcome Trust. (2003). *Economic analysis of scientific research publishing.* Cambridgeshire. Retrieved from http://www.wellcome.ac.uk/en/images/ SciResPublishing2_7445.pdf

Wellen, R. (2004). Taking on commercial scholarly journals: Reflections on the “open access” movement. *Journal of Academic Ethics, 2*(1), 101–118. http://doi.org/10.1023/B:JAET.0000039010.14325.3d

White, H. D. (2001). Authors as citers over time. *Journal of the Association for Information Science and Technology*, *52*(2), 87–108.

Whiteside, E., & Hardin, M. (2011). Women (not) watching women: Leisure time, television, and implications for televised coverage of women’s sports. *Communication, Culture & Critique*, *4*(2), 122–143. http://doi.org/10.1111/j.1753-9137.2011.01098.x

Willinsky, J. (2006). *The access principle*. Cambridge Massachusetts: MIT Press.

Wingfield, N., Isaac, M., & Benner, K. (2016, November 14). Google and Facebook take aim at fake news sites. *New York Times*. Retrieved from http://www.nytimes.com/2016/11/15/technology/google-will-ban-websites-that-host-fake-news-from-using-its-ad-service.html?_r=0

Wooldridge, J. M. (2013). Introductory econometrics: a modern approach. Mason, Ohio: South-Western Cengage Learning.

Zeigermann, L. (n.d.). Opencagegeo: Stata Module for Forward and Reverse Geocoding. Retrieved from http://fmwww.bc.edu/RePEc/bocode/o/opencagegeo.pdf

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