

**SOCIAL GATEKEEPING, THE SERENDIPITOUS TIE AND DISCOVERY:  
AUTHORS CONNECTING READERS TO BOOKS THROUGH SOCIAL MEDIA  
OUTREACH**

By

Bruce Fulton

---

Copyright © Bruce Fulton 2013

A Dissertation Submitted to the Faculty of the

SCHOOL OF INFORMATION RESOURCES AND LIBRARY SCIENCE

In partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2013

THE UNIVERSITY OF ARIZONA  
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Bruce D. Fulton, titled Social Gatekeeping, the Serendipitous Tie and Discovery: Authors Connecting Readers to Books through Social Media Outreach and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

\_\_\_\_\_ Date: 7/19/2013  
Jana Bradley

\_\_\_\_\_ Date: 7/19/2013  
Martin Frické

\_\_\_\_\_ Date: 7/19/2013  
P. Bryan Heidorn

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

\_\_\_\_\_ Date: 7/19/2013  
Dissertation Director: Jana Bradley

\_\_\_\_\_ Date: 7/19/2013  
Dissertation Director: Martin Frické

### STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Bruce Fulton

## ACKNOWLEDGEMENTS

This research project and completion of my PhD would not have been possible without the support, guidance and love of many individuals. A special thanks of gratitude and appreciation goes to my committee co-chair and dissertation advisor Jana Bradley, who went above and beyond to help make this work the best it could be.

Thanks and appreciation also go to dissertation committee co-chair Martin Frické and to committee advisor P. Bryan Heidorn for their thoughtful consideration and evaluation of this research throughout the revision process and final defense.

Thanks and appreciation are due for the additional help and guidance from minor advisor Stephen Rains and comprehensive committee advisor Patricia Montiel-Overall, who were invaluable in helping me get to the dissertation stage.

A special note of thanks goes to the staff of the School of Information Resources and Library Science at the University of Arizona who guided and smoothed my journey through the administrative tasks necessary to complete the degree process, and to the full faculty who provided encouragement, support and valuable advice and counseling throughout my studies.

Finally, thanks and love go to my wife, Lorrane, whose patience, love and faith throughout the difficult and challenging dissertation journey gave me the strength, inspiration and confidence to finish the task.

## TABLE OF CONTENTS

LIST OF TABLES .....	11
LIST OF FIGURES .....	13
LIST OF EQUATIONS .....	14
ABSTRACT .....	15
CHAPTER 1 - INTRODUCTION .....	17
Problem Statement – The Discoverability Problem .....	19
Background of the Problem – Current Understanding .....	23
Purpose .....	27
Theoretical Framework.....	28
The Publishing Chain .....	28
Gatekeeping Theory.....	29
Gatekeeping: Extending the Publishing Chain Framework .....	32
Information Diffusion in Social Networks.....	35
Toward a Theory of Social Gatekeeping.....	39
Identification and Location of the Message as a Unit of Analysis .....	44
The Serendipitous Tie.....	46
Measuring Social Gatekeeping .....	51
Social Gatekeeping as Strategy.....	55
Research Questions.....	58
Nature of the Study .....	64
Assumptions .....	66
Limitations and Significance of the Study .....	68

**TABLE OF CONTENTS – Continued**

Definition of Terms .....	71
CHAPTER 2 - REVIEW OF THE LITERATURE .....	77
History and Definition of eBook.....	77
Publishing and the Publishing Chain.....	81
Gatekeeping.....	85
Discovery: Serendipity and Browse .....	94
Empirical Support for Social Gatekeeping and the Serendipitous Tie.....	98
Social Gatekeeping, Sales and the Review .....	104
Popular Conceptions of Gatekeeping .....	108
Multiple Regression.....	116
Search Engine Count as Dependent Variable .....	119
Amazon Sales Rank as Dependent Variable.....	123
CHAPTER 3 - METHODOLOGY .....	125
Introduction to the Methodology .....	125
Explanation of Data Collection Techniques.....	126
Technology Environment .....	129
Description and Collection of the Random and Popular Samples.....	130
Random Sample .....	130
Popular Sample .....	133
Preparation of the Datasets.....	135
Dependent Variable Weekly Data Collection.....	136
Strategy and Approach .....	136
Dependent Variables - Weekly Data Collection.....	138
Collection and Classification of Independent (Predictor) Variables .....	144

**TABLE OF CONTENTS – Continued**

Strategy and Approach .....	144
Collection of the Independent variables .....	145
Classification of Authors, Titles and Publishers .....	149
Data Analysis and Statistical Approach .....	152
Strategy and Approach .....	152
Statistical Setup and Preparation.....	152
Independent Variable Frequency Tables .....	157
Contamination of the Dependent Variable.....	158
Ethical Considerations .....	162
CHAPTER 4 - RESULTS .....	163
Phase I – Description, Frequencies and Central Tendencies of the Random and Popular Samples.....	163
Research Questions and Hypotheses – Phase I.....	163
Language .....	165
Reprint Jungle .....	166
Other Adjustments, Initial .....	166
Self-Published Books.....	167
General Availability.....	168
Print Availability.....	169
Miscellaneous Attributes .....	170
Final Adjustments to the Samples.....	173
Sales – Random Sample .....	174

**TABLE OF CONTENTS – Continued**

Sales as a Function of Social Media Outreach – Random Sample .....	174
Price Manipulation – Random Sample .....	175
Pricing – Random Sample .....	176
Phase II – Quantitative Analyses .....	178
Purpose .....	178
Primary Regression Models .....	178
Regressions computed on the Random Sample .....	182
Summary – Random Sample Regressions.....	189
Regressions Computed on the Popular Sample .....	191
Summary – Popular Sample Regressions.....	198
Ad Hoc Regressions.....	200
Correlations .....	206
Analysis of Reviews .....	209
Research Questions and Hypotheses – Phase II.....	212
Phase III - Title Review .....	216
Purpose .....	216
Research Questions – Phase III.....	216
Review of Selected Titles from the Random Sample.....	217
Briefly Noted from the Popular Sample.....	230
Summary.....	233

**TABLE OF CONTENTS – Continued**

CHAPTER 5 - DISCUSSION .....	236
Introduction .....	236
Summary of the Purposes of the Research, Theoretical Framework and Methodology .....	237
Summary of the Results with Conclusions.....	244
Phase I Research - Discussion.....	244
Phase II Research - Discussion .....	252
Phase III Research - Discussion.....	263
Significance of the Research .....	266
Implications for Future Research and Recommendations.....	267
APPENDIX A - INDIVIDUAL REGRESSION RESULTS – COMPLETE .....	270
DV = Google search engine counts on ASIN, Random Sample.....	270
DV = Google search engine counts on ASIN, Popular Sample.....	273
DV = Google search engine counts, ASIN in Blog Pages, Random Sample.....	275
DV = Google search engine counts, ASIN in Blog Pages, Popular Sample.....	276
DV = Google search engine counts on Author - Title, Random Sample .....	278
DV = Google search engine counts on Author - Title, Popular Sample .....	279
DV = Bing search engine counts on ASIN, Random Sample.....	281
DV = Bing search engine counts on ASIN, Popular Sample .....	283
DV = Bing search engine counts on Author – Title, Random Sample.....	285
DV = Bing search engine counts on Author – Title, Popular Sample.....	287
DV = Amazon Sales, Random Sample.....	289
DV = Amazon Sales, Popular Sample.....	291

**TABLE OF CONTENTS – *Continued***

DV = Amazon Review Count, Random Sample.....	293
DV = Amazon Review Count, Popular Sample.....	295
APPENDIX B – RANDOM SAMPLE TITLES.....	297
APPENDIX C – POPULAR SAMPLE TITLES.....	348
WORKS CITED.....	372

## LIST OF TABLES

Table 1 - Author/Title Normalization.....	139
Table 2 - Data Collection, Dependent Variables .....	142
Table 3 - Data Collection, Independent Variables.....	148
Table 4 - Distribution of the predictor variables from the Random Sample.....	157
Table 5 - Distribution of the predictor variables for the Popular Sample .....	157
Table 6 - Foreign Language Counts .....	165
Table 7 - Pricing, Self-published Books .....	176
Table 8 - Pricing, Traditionally Published Books .....	177
Table 9 - Google ASIN Betas, Random Sample .....	182
Table 10 - Bing ASIN Betas, Random Sample.....	185
Table 11 - Bing A/T Betas, Random Sample.....	186
Table 12 - Amazon Sales Betas, Random Sample .....	187
Table 13 - Amazon Reviews Betas, Random Sample.....	188
Table 14 - Summary of Significant Predictors, Random Sample.....	189
Table 15 - Google ASIN Betas, Popular Sample .....	191
Table 16 - Google A/T Betas, Popular Sample.....	193
Table 17 - Bing ASIN Betas, Popular Sample.....	194

**LIST OF TABLES – *Continued***

Table 18 - Bing A/T, Popular Sample .....	195
Table 19 - Amazon Sales Betas, Popular Sample.....	196
Table 20 - Amazon Reviews Betas, Popular Sample .....	197
Table 21 - Summary of Significant Predictors, Popular Sample.....	198
Table 22 - Print Version Betas, Random Sample .....	200
Table 23 - Fiction Betas, Random Sample.....	203
Table 24 - Non-Fiction Betas, Random Sample .....	203
Table 25 - DV Correlations, Random Sample .....	207
Table 26 - Summary of Betas, Random Sample.....	254
Table 27 - Summary of Betas, Popular Sample.....	255

**LIST OF FIGURES**

Figure 1 – The Traditional Publishing Chain.....	33
Figure 2 - Extending the Publishing Chain.....	41
Figure 3- Author Web Presence.....	158
Figure 4 - Plot, Rank vs Review, Random Sample .....	210
Figure 5 - Plot, Rank vs. Review, Popular Sample .....	211

**LIST OF EQUATIONS**

Equation 1 – Sales as a Function of Sales Rank .....	153
Equation 2 – Moving Weighted Average Formula .....	154
Equation 3 - Moving Weighted Average Example .....	154
Equation 4 - DV Contamination Formula 1 .....	159
Equation 5 - DV Contamination Equation 2.....	159
Equation 6 - Sales Rank as a Function of Review Count.....	209

## ABSTRACT

In 2011, over 1.5 million new book titles were published in the United States, a 400% increase in just five years compared to 2006. In the same time period, the market share for eBooks increased dramatically and now comprises 20% or more of sales from many of the biggest publishing companies. This hyper-abundance of titles in an increasingly heterogeneous market place has made it difficult for consumers to connect to books they might want to read.

This is the discovery problem. It is compounded by the continuing decline of traditional gatekeepers and sources of discovery such as mass media reviews and advertising, as well as the decline of traditional bookstores where people often find books through browse. Authors and publishers therefore have turned to social media to spread the word about their titles. Social gatekeeping, an extension of traditional gatekeeping theory, is proposed as the framework for understanding how author participation in social networks initiates a flow of the diffusion of information over the web and other computer mediated communication channels, and through individuals and social networks to potential readers. Serendipitous browse and discovery is a key strategy for readers to find titles of interest, and the serendipitous tie is proposed as a social mechanism through which individuals discover new titles and bring it back to their social networks to share.

To explore these concepts, a random sample of new eBook titles published during the first week of April, 2012 was generated and analyzed in three phases. The first phase of research classified books and authors according to facets such as traditional or self-

published, use of social media and other factors. The second phase used multiple regression to establish an association between the use of social media by authors and a title's sales and presence on the Web. The third phase reviewed selected titles for new approaches to social media use and evidence of the serendipitous tie. The results are consistent with the hypothesis that author web presence predicts discoverability and sales.

## CHAPTER 1 - INTRODUCTION

The number of book titles published annually in the United States has increased dramatically in recent years, from an estimated 296,352 titles in 2006 to an estimated 1,532,623 in 2011, as measured by new ISBN numbers issued (Bowker, 2012a). These figures don't include an unknown number of additional print and digital book titles for which no standard ISBN number was sought or issued. A breakdown of production according to the same Bowker data reveals that title output of traditionally published books<sup>1</sup> from the major publishing conglomerates has increased somewhat or remained flat, while non-traditionally published books including books from small and niche publishers as well as self-published titles and a burgeoning reprint market account for the lion's share of the increases.

Today, anyone who seeks publication can be published, thanks largely to technology and the ubiquity of the Web and other communication channels, which have reduced or eliminated traditional barriers to the publishing process. Inexpensive desktop publishing programs available on personal computers have lowered the cost and complexity of typesetting and pre-publication preparation. Digital printing technologies now compete in cost with mechanical offset printing and are easier to use. Digital printing also facilitates print-on-demand (POD) business models that reduce or eliminate investment in print runs and warehousing. And finally, electronic text such as eBooks and

---

<sup>1</sup> Usage note: unless otherwise distinguished, the term "book" used in this document refers generically to both print and digital versions of books. The term "eBook" refers specifically to a digital version of a book.

eBook applications eliminates printing, warehousing and transportation costs entirely.

For a growing number of publishers and authors today, many books are only published in digital format, without a print analog ever being produced.

Digital technologies and the Web have also spawned a plethora of venues to find, evaluate and retrieve print and digital text. While traditional gatekeepers such as publishers, literary agents, booksellers and professional reviewers still exert considerable influence on book production and sales, individual authors and small/niche publishers as well as the major players can find alternative channels through which to produce content and connect readers with it. Today's emerging publishing landscape is heterogeneous and hyper-abundant (Bradley, Fulton, Helm & Pittner, 2011; Lichtenberg, 2011). There are more choices of books to read than the reader could possibly discover let alone evaluate, and they are distributed through a galaxy of distributors, most of them online, and many of them obscure.

### **Problem Statement – The Discoverability Problem**

The problem of resource discovery in a network of distributed resources has been recognized since the early days of the Internet. As the number of users and resources began to increase, methods were developed to help users navigate through an exponentially growing number of links and cross-references. Archie, an early pre-Web Internet indexing application, was developed specifically to address what its developers referred to as the Resource Discovery Problem (Deutsch, 1991). Similar problems began to emerge as computer databases became larger, both in size and number. The Text Database Discovery Problem describes the challenges associated with selecting appropriate databases of text-based documents (Gravano et al, 1993). Knowledge Discovery refers to data-mining, the process of discovering non-trivial patterns and relationships hidden in large datasets (Ziarko, 1995).

By 1998, early Web and database indexing approaches began to collapse as millions of new Web pages on hundreds of thousands of network nodes came on-line. Then, it was estimated that a billion Web pages would be published by 2000, prompting the development of the Google page rank system of relevancy ratings designed to filter and order the large number of otherwise unmanageable search returns (Brin and Page, 1998).

The problem for the reader today is precisely this hyper-abundance of choice exacerbated by the number of available sources of discovery and acquisition. Increasingly, these sources are Web-based even as traditional sources such as print

reviews in newspapers and magazines and physical bookstores continue to decline.

Readers ask, “How can I find and acquire the books I desire?”

In order to connect to books that satisfy their needs and desires, readers must find ways to navigate their way through the glut of titles and distribution sources. Readers are no longer limited to (or by) a small number of known traditional gatekeepers – individuals, companies or organizations that filter, select, organize, manage and promote reading options. Today’s readers choose not only from a hyper-abundant selection of titles but through a large and growing pool of alternative and non-traditional gatekeepers – new sources of discovery, evaluation and acquisition – some of which are human and some of which are now automated discovery processes and software applications. Readers may continue to depend on the traditional gatekeeping processes of publishing mainstream in order to discover books but in doing so may miss alternative choices that new models of discovery could reveal.

The problem for the author mirrors that of the reader. “How does the reader find my book?” The author must figure out a way to connect reader to the book. When there were fewer choices, there were fewer paths to discovery. As the number of titles and sources grows, the number of potential paths to discovery grows exponentially. The author’s challenge is to try to assure that the path leading to the book is clear for the reader who seeks it. The author therefore also now faces new gatekeeping challenges and choices. Some authors continue to rely on the gatekeepers of traditional publishing, for those able to negotiate a mainstream publishing contract. With the exception of a limited amount of vanity publishing, a mainstream publishing contract was the only choice for

authors in the mid- to late-20<sup>th</sup> century. As digital publishing technologies have become widespread, authors may also now connect readers to books through newly emerging information gates and their gatekeepers, or through some combination of traditional and emerging marketing models of the last decade.

Discoverability has become a popular theme and issue of concern in publishing trade circles beginning in 2008. The adoption of the term in that context coincides with the beginning of a rapid rise in the popularity of eBooks and a simultaneous and rapid rise in the adoption of mobile computing devices and so-called app stores, where software application authors vie for downloads in much the same way that book authors vie for readers. Google analytics show that the term first shot to prominence (in contexts other than Bluetooth discovery technology) in 2007, with a peak in 2008, and continued strong ongoing interest continuing to trend upward (<http://www.google.com/trends/explore#q=discoverability%20-bluetooth>), retrieved 10/15/2012). Google aligns several points on the trend chart with articles on the discoverability of software applications, primarily on mobile computing devices, in app stores.

As with many terms entering the vernacular, the term *discoverability* used in the context of publishing lacks concise definitional rigor and perhaps means different things to different people or communities of practice. There is a coalescence of usage in eBook publishing circles, however, that typically relates the problem of discoverability to a solution subset of emerging marketing techniques that converge on digital discovery through online channels. Industry trade association Digital Book World's 2012

Conference on Marketing and Discoverability exemplifies this approach (<http://tinyurl.com/8g62tj5>). Thus, the book trade's solutions to discoverability and eBook discoverability in particular focus increasingly on alternative online awareness and marketing, as contrasted with traditional non-virtual marketing techniques such as print and mass media advertising, co-op bookseller displays, and physical / face to face promotions. This may stem at least in part from the growing number of digital only or digital first literary works where access to information about the work and acquisition of the work occurs almost exclusively online.

This research examines one aspect of the discoverability problem, the use of computer-mediated networks in the context of today's heterogeneous hyper-abundant book market, and seeks to establish whether and to what extent social media is being used effectively by authors to connect readers to their books. In particular, the research more narrowly focuses on eBooks as an emergent form of literary production that by its very nature can only be acquired and read through some form of computer-mediated communication medium.

## **Background of the Problem – Current Understanding**

After a decade or more of largely failed predictions concerning the ascendancy of the eBook and a corresponding decline in production of print media, 2011 finally saw eBook and digital reading device sales rise dramatically compared to traditionally published print books, sales of which remained flat or in decline among most major publishers (Flood, 2011). Although the majority of book sales remain firmly on the print side, eBooks comprised up to 20% or more of sales among the traditional publishing conglomerates in the US and UK in 2011, up from single digits just a year earlier. Some industry analysts now predict that eBooks will eventually account for as much as 50% of sales for the mainstream trade press, perhaps within a decade or less. Publishers Weekly quotes Hachette's CEO as saying, "Digital sales will represent close to 30% of Hachette Book Group's net revenue in 2012, and could reach 45%–50% for us by 2015." (Publishers Weekly, Dec 30, 2011)

The extent of the emerging heterogeneous book market is largely unknown. Bookscan, a division of the Neilson Company that tracks print book sales, has only recently begun limited tracking of eBook sales primarily through a few mainstream vendors. Since few eBooks are sold through point of sale collection devices, what is known about eBook sales figures currently is what publishers choose to report and what a few online merchants are tracking via online sales monitoring and data-sharing with Bookscan. For the most part, this covers only mainstream publishers and major channels to market and does not include author direct –to-consumer sales and small / niche channels to market (Charman-Anderson, 2013).

Although the library and information science literature has a long tradition of research on search and browse in the context of information organization and the use of library tools such as the OPAC, little is found in the library literature on eBook discovery per se. Few libraries currently acquire and lend eBooks directly; instead, most libraries host online portals that farm management of eBook lending out to commercial ventures such as Overdrive and directly to Amazon itself. Title selection is limited, and many major publishers currently decline to allow their books to be used in library lending programs (Greenfield, 2012). Libraries also collect relatively few self-published titles (Dilevko & Dali, 2006), which now comprise a significant portion of the output of both print and digital literary production (Bradley et al, 2012).

Most of what is currently known about eBook discovery specifically, and book discovery in today's current heterogeneous marketplace in general, comes from surveys and studies done by for-profit trade associations, social network providers, and a few scholarly survey sources such as the Pew Foundation media studies surveys.

Goodreads.com CEO Otis Chandler delivered a keynote address entitled "How People Discover Books Online" at the 2012 Tools of Change Conference based on data from the book-focused social network's several million users (Chandler, 2012). Some statistics from that research indicated that of readers who discovered books (print and digital) on the Goodreads site itself, 19% discovered books through site search, 19% from the registration process, 13% from recommendations and the Goodreads socially driven recommendation engine, 9% from friend updates, 9% from genre browsing, 8% from

author or series pages, 7% from various lists posted on the site, 6% from mobile applications, 2% from advertised giveaways, and a small number of “other.”

Chandler’s went on to say that of people who discovered books in ways other than on the Goodreads site, 96% discovered books by known author, 79% from offline or out of network friends, 59% from bookstores, 54% from Amazon, 54% from libraries, and fewer numbers from browsing book sites, newspapers, author’s website, radio, TV, Facebook, Twitter and publishers’ websites. The data also indicated significant spikes in book activity following mass media events such as reviews or interviews in major media outlets such as NPR or the New York Times, indicating that mass media gatekeepers still exert a significant impact on awareness. These results may not be typical, since Goodreads social network members are generally considered avid readers whose habits, by virtue of their participation in the book-focused social network, may not mirror that of the general public.

Verso Digital’s 2011 Survey of Book Buying Behavior surveyed internet users in a poll statistically weighted to mirror the US population as a whole (Verso, 2012). The survey estimated a population of 70 million avid book-buyers purchasing 10 or more books per year including eBooks. The discoverability breakdown for books in all formats is as follows: 49% personal recommendations, 30.8% bookstore staff recommendations, 24.4% advertising, 21.6% search engine, 18.9% book reviews, 16% online algorithm, 15.5% library visit, 12.1% blogs and 11.8% social networks. Unfortunately, the survey does not break out discovery practices according to print versus digital format.

There is evidence, however, that discoverability patterns vary between the general public and the eBook-reading public. The Pew Internet and American Life survey on e-reading conducted in December of 2011 reported that 81% of e-reader and tablet owners got recommendations from family members, friends, and co-workers vs. 64% for the general population. For online bookstores or other Web sites, the figures were 56% for e-reader and tablet owners versus only 28% for others, indicating that online presence may be of greater importance for discovery of eBooks than of book titles in general (Rainie et al, 2012).

These numbers suggest at least two important lines of inquiry relating to discovery. First, friends and family, through one means or another, remain important for the discovery of eBook titles, implying that diffusion of information through social networks, either online or face to face, is a key strategy for authors and publishers who want to make new titles discoverable. Second, eBooks represent a new and quickly growing segment of book production and literary output. Discovery of eBooks through online means is a relatively recent phenomenon that is only poorly understood compared to discovery of print books through traditional means.

**Purpose**

The purposes of this research are to describe the use of social media by eBook authors and to determine the extent to which such use increases discoverability and readership. The research is conducted through the lens of social gatekeeping, described further in the sections that follow, which provides a framework for understanding how author participation in social networks and other computer mediated communication channels initiates a flow of the diffusion of information over the web and other computer mediated communication channels and through individuals and social networks to potential readers.

## **Theoretical Framework**

### *The Publishing Chain*

A number of intermediaries stand between the author and the reader; this network of individuals, companies and processes has been called the publishing chain (Thompson 2010). In traditional mainstream publishing as practiced today, the author first typically signs with a literary agent designated to negotiate on the author's behalf with a publisher. The agent often works with the author to polish a rough manuscript or book proposal and then shops the resulting work (or work in progress) to a publisher.

Once a publisher decides to accept a book for publication, value is added to the raw manuscript including art, copyediting, typesetting, layout and then printing. Printed books are shipped to a distributor that provides warehousing and transportation, and from there, the book continues its physical journey through wholesalers and various bookseller outlets toward its final destination, namely the reader along with institutions such as libraries. Traditionally published eBooks wind their way through a similar, albeit virtual, network of distributors and intermediaries.

The publisher also makes metadata, information about a book, available to aid in discovery – facts such as author and title, length, a description or abstract, pre-release reviews and chapter lists, and so on. It is at this point that the publishing chain forks into two channels. While the book is routed through distribution channels that ultimately provide print or digital access, information about the book is distributed through communication channels to ensure that readers become aware of the book and its salient

characteristics. Marketing, including advertising, review copies, co-op merchandising with booksellers (such as featured display tables in book stores for print books) and other promotions, are some ways readers come to discover books by way of the mainstream model.

Professional reviews are another way readers discover books. Thus, news information gatekeepers including reviewers, reporters, editors and publishers (or producers) provide another independent information channel, courted and solicited by publishers as a complement to mass media marketing.

Finally, discovery is aided by information distributed via word of mouth. Individuals prompted by initial discovery through marketing and news channels spread information and recommendations to family, friends and acquaintances through social networks, which in the pre-Web world were largely face to face.

### *Gatekeeping Theory*

Gatekeeping as a term and descriptive framework was introduced in 1943 by Kurt Lewin (Lewin, 1943; Lewin, 1947), who was researching why people eat what they eat and how food comes to the dinner table. Lewin first introduced channel theory, saying that food comes to the table through channels, such as from the grocery store or from the garden, and that food moves in steps through the channels. Further, he notes that food doesn't move itself through the channel and that at various steps along the way, decisions about whether and what to move along are made by individuals he calls gatekeepers. Each channel may have multiple gatekeepers; for example there are decisions a grocery

store makes about what to carry, decisions the shopper makes about what to select, decisions about how and whether to pay for it, and decisions about how to prepare it. Each of these constitutes a gate, and the gatekeepers collectively determine what ends up on the table. Then, of course, the individual makes the decision about whether and what to actually consume.

Lewin was primarily interested in how people might come to change food habits in the context of scarcity, rising prices and rationing in the early years of World War II. Lewin suggested that changes, if they are to be made, are made by gatekeepers at critical decision points in the food channels. Lewin suggests that factors affecting the gatekeeper's decisions – the gatekeeper's "food ideology" – include cognitive structures and beliefs (such as beliefs about health and nutrition norms), motivation (money, health, taste, status) and conflict (competing values). Lewin created study groups and evaluated decision-making strategies effecting change in food choice; he concluded, based on experimental evidence, that influencing gatekeepers at various decision points within the food to table channels resulted in faster more effective changes in food habits compared to other strategies such as lecture and request.

Gatekeeping has since appeared as a theory, factor or framework in studies across a number of disciplines including political science, sociology, law, health sciences, communication, journalism and library science (Barzilai-Nahon, 2008; Barzilai-Nahon, 2009). The basic understanding of gatekeeping has changed little since Lewin first proposed it, but the disciplines have contextualized it in different ways. Two concepts are especially useful to this discussion.

Editorial gatekeeping posits the gatekeeping framework as a theory of information filtered out. Editorial gatekeeping focuses on an article, paper, report or news item as a message that is either rejected, or accepted and packaged for the reader. Lewin (1947) himself first suggested that news flowing through a communications channel was subjected to gates controlled either by a gatekeeper or an impartial rule. White (1964) was the first to formalize this conceptualization within the disciplines of communication and journalism with a study of a news editor and the great numbers of articles flowing through the editor's desk via reporters and the wire services. The stories either are rejected by the gatekeeping process of editorial decision-making (most of them) or pass through the editor's gate and into publication for the reader (only a few of them). According to White, the factors influencing a gatekeeper's decision to reject or publish a story reflect individual differences in psychology, culture and experience.

In contrast, other disciplines, especially including Library and Information Science, focus on the gatekeeper as an information intermediary and posit that gatekeeping is a process of information filtered in. Libraries, for example, rarely try to acquire every book and resource available on a particular topic. Their goal is to create a collection of resources from a universe of titles that best represents diversity of opinion and that reflects the interests and culture of the community they serve. The librarian's goal therefore is not to filter out books that fail to meet criteria established for selection but to select in the best collection of books and other resources that do. Many resources may meet basic selection criteria, but with limited time and funds, only a few can be acquired. In this context, the gatekeeper serves an important role in discovery (Sturges,

2001; Su & Contractor, 2011) and cultural preservation (Joyce, 1998; Lim, 1995). This is primarily a construct of information filtered in.

Another term referring to the person in charge of selection and management of a collection, usually in contemplation of the cultural and educational needs of particular communities of users, is *curator*, and for that reason, the term *curatorial gatekeeping* is applied here to this particular sense of gatekeeping. A case can be made that editorial and curatorial gatekeeping are two sides of the same coin. The editor may be viewed as filtering in stories that meet reader expectations and libraries may be obliged to filter out material unsuitable for their audience. Arguably, the distinctions are a matter of perspective and depend upon who may have something to gain or lose as a result of the gatekeepers' decisions.

#### *Gatekeeping: Extending the Publishing Chain Framework*

The flowchart shown in Figure 1 (next page) is based on and extends a classic conceptualization of the publishing chain as presented by Thompson (2010, p. 16). It is expanded to illustrate the processes essential to understanding the implications of gatekeeping theory on key publishing chain concepts. First, the new chart explicitly illustrates the diverging (and then re-converging) information flows of the book as object (noted in green) and the information about the book (metadata, noted in red). This holds clear the difference between the discovery process through metadata, and access plus retrieval of the physical (or digital) object, which generally occurs as a result of, but not necessarily in conjunction with, discovery.

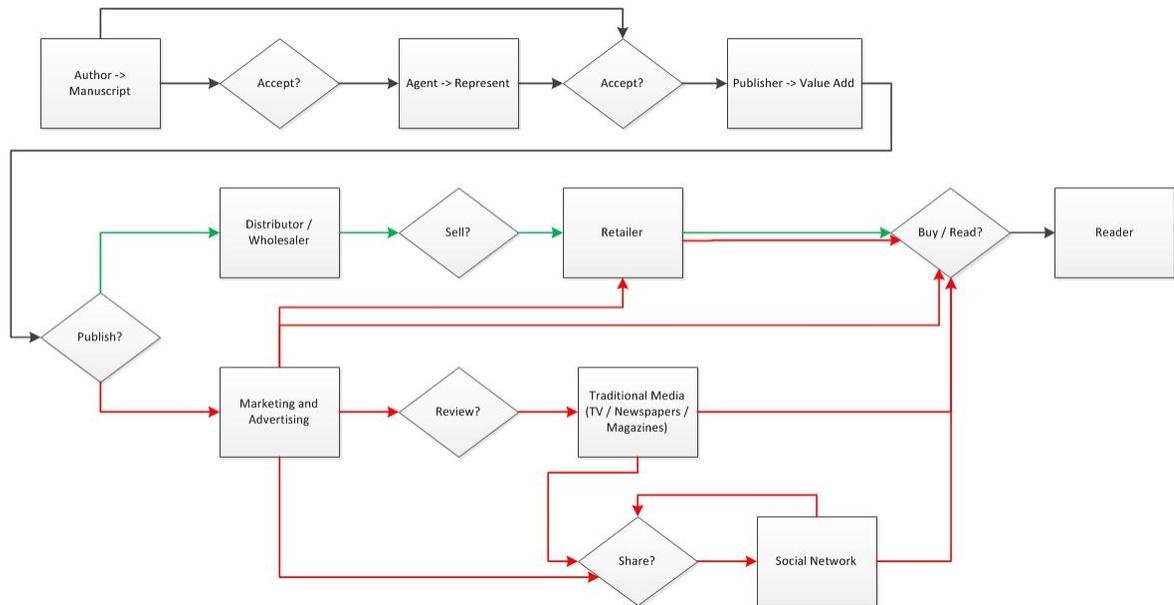


Figure 1 – The Traditional Publishing Chain

The second feature differentiating this chart from its predecessors is that decision points representing gatekeeper decisions are explicitly indicated rather than implied. For example, a literary agent may decide to accept an author's book for representation or not. The diamond-shaped figure "accept?" between the author entity and the agent entity is the decision point. If the agent accepts the manuscript, it passes through to the agent. Similarly, the publisher stands between the agent and the publishing process. The decision points in the flow of information through various channels in the book publishing chain are information gates, and the individual or entity making a decision about whether information continues to flow through the information channel at that point is a gatekeeper.

Modern traditional publishing, which is almost all modern publishing prior to the beginning of the 21<sup>st</sup> century, falls under the gatekeeping framework by virtue of the gatekeeping roles of the editor and publisher (Cosser, 1975), and also of the booksellers who make independent decisions about what to stock. Through the end of the 20<sup>th</sup> century, prior to the development of Web and digital technologies described earlier in this chapter, authors had little choice but to deal directly with the gatekeepers of the book publishing chain. Since publishers reject the great majority of manuscripts they receive for consideration, the process aligns with the information filtered out perspective of communication and journalism, at least from the point of view of the aspiring author.

A few authors of the mid- to late-twentieth century attempted self-publishing, many of them through vanity publishers, which arranged costly offset printing at the author's expense. But the vanity publisher had no way of bypassing traditional publishing's gatekeepers, which effectively filtered out from public purview titles that did not meet their standards or needs. Without access to the publisher's established supply chain, self-published authors were limited in their ability to reach an audience, either with information about the book (metadata) or with the book itself. While reported production of vanity press titles sometimes appeared comparable to output from the mainstream publishers, few copies actually entered the retail channel (Sullivan, 1958; Milliot & Coffey, 2010) or were collected by libraries (DiLevko & Dali, 2006), with most printed copies sitting unsold in the author's garage or basement. Sales occurred almost exclusively through the author's professional or social network, which pre-Web was almost all face to face.

### *Information Diffusion in Social Networks*

Market research cited previously shows that family and friends are important sources for the discovery of new titles, with numbers and percentages varying depending on the individual survey. This finding indicates that social networks play an important role in book discovery, and it also suggests two related questions: 1) how did the friend or family member find out about the book in the first place, and 2) how does information about books flow to, and then among and between, individuals in the different kinds of online and face to face social networks?

Early research on the flow of information from a mass media source to the general public began in the 1940s with the theory of the two-step flow of communications. Proposed and developed by Lazarsfeld et al (1944) and further developed by Lazarsfeld and Katz (1955), the two-step theory proposed that "ideas often flow from radio and print to opinion leaders and from these to the less active sections of the population" (Katz, 1957 pg. 61). Studied originally in the context of choice-making in a political campaign, the theory was among the first to challenge the prevailing opinions on the power of mass media to directly influence an audience (the so-called hypodermic needle model). It suggested that opinion leaders served as an intermediary between the information pushed by mass media and its ultimate consumption by the population at large.

The predictions of the theory were generally sustained by additional studies in the 1940s and 1950s, with greater attention to the nature of the interpersonal factors affecting influence and chains of influence extending beyond the dyad. In reviewing confirmatory studies, Katz (1957) called attention to the importance of studying diffusion over time

and suggested that studies might be developed that incorporated the elements of a specific item, diffusion over time and the social structure of an entire community in order to isolate the roles these factors play in the flow of opinion change.

Rogers expanded on the concept of the opinion leader and earlier diffusion research to develop the theory of the Diffusion of Innovations (1962, 1995, 2003). As presented, the theory is an overarching framework drawn from a number of disciplines that explains how ideas and technologies are adopted over time within a social network context. Taken as a whole, the theory proposes that differences in innovation adoption are a function of product characteristics (e.g. how much of an advantage the innovation has over existing technology, how easily it may be tested, etc.) and individual differences along a dimension of resistance or readiness to accept and implement new ideas. According to Rogers, knowledge is spread through information channels including mass media and social channels through opinion leaders, and acceptance takes place over time in stages by individuals with varying degrees of acceptance tendencies, identified as innovators and early adopters to early and late stage majorities and finally to the laggards. Rogers positions *knowledge* and *persuasion* as the first two stages of the diffusion/adoption process, and in the fourth edition of Diffusion of Innovation (1995), he notes that:

“Diffusion and Adoption Gatekeeping is controlling the flow of messages through a communication channel. One of the most crucial decisions in the entire innovation-development process is the decision to begin diffusing an innovation to potential adopters.” (p. 148).

An effective strategy suggested by both the Diffusion of Innovations and Two-step Flow of Communications theories, is to influence the opinion leaders, change agents and early adopters who spread information that leads to adoption of ideas and new innovation.

Social network analysis is the primary tool by which the specifics of the diffusion of information through social networks are modeled. Broadly speaking, social network analysis, derived from the concepts and mathematics of network analysis, positions individuals as nodes on a network, with connections between and among them defining relationship characteristics (Borgatti et al, 2009). Social networks, either face to face or computer mediated, can be modeled or experimentally described and measured by types, numbers and characteristics of the nodes and connections. Social network analysis has been applied to the study of the flow of information through social networks, and two variables come into play more often than not.

Heterophily is the “degree to which pairs of individuals who interact are different in certain attributes (Rogers, 2003, pg 306).” Heterophily and its antonym homophily, which is the degree to which individuals are alike, play interesting and complementary roles in the diffusion of information and adoption of innovation in a network. Rogers observed that adoption influence was greatest among those who were the most homophilous, that is, you are most likely to be influenced to adopt an innovation or opinion by someone similar to you in attitudes, beliefs and other traits. On the other hand, the more two individuals are alike, the greater the likelihood is that both individuals will already know certain information or have adopted certain new innovations. According to

Rogers, the strongest persuasion may come when the only difference in a homophilous dyad is the information or innovation itself, but homophilous dyads are a poor source for new information. Effective diffusion therefore requires some degree of heterophily.

Why this might be so was explained by Granovetter (1973) who proposed that new information often comes to an individual from those in one's network who are socially distant, or so called weak social ties, as opposed to those who are close in one's network, or strong ties. The strength of a tie is measured by how often and to what degree two individuals communicate. There are two factors that seem to explain this. The first is that weak ties tend to be heterophilous and therefore diverge more widely from strong ties in knowledge, connections, attitudes and behaviors. The second, and perhaps more important factor is that weak ties tend to be much more numerous than strong ties, that is, one has a much larger set of social acquaintances than they do close friends and family. Particularly in today's large and extensive computer mediated social networks where "friends" may number in the hundreds, network analysis has demonstrated that the sheer numbers of weak tie links account for most of the new information that diffuses through a network, compared to strong tie communicators (Bakshy et al, 2012).

Even so, not all information diffusion occurs through ties that can be categorized as strong or weak, and research also shows that information can become isolated within individual communities (Bakshy et al, 2012; Onnela et al, 2007) and may not continue to spread without mechanisms for information flow beyond that explained by the weak tie.

### *Toward a Theory of Social Gatekeeping*

While traditional gatekeeping theory has proven a good fit for mass media communication channels generally, and the mainstream publishing industry specifically as applied here, it has seen little application in research as a framework for the flow of information among and between individuals. Similarly, diffusion of innovations theories generally have tended to focus more on aspects of individual differences in adoption readiness and inclination and less on the specifics of the gatekeeping processes that control the initial flows of information between individuals and among social groups and communities. This section extends gatekeeping theory in order to address these gaps of current understanding

The decision by a reader to engage with a book is dependent on at least two things. First, the reader must discover the book, that is, become aware that it exists. Second, the reader must learn enough about the book to determine that reading it will have some probability of satisfying a need, either for information or entertainment or some other combination of factors. The information that could help a reader decide to acquire the book might include information about the content, price, length, available format and other information. Dust jacket blurbs, reviews, blogs and other kinds of third-party evaluations may also play a role in persuading the reader to act, and research has shown that these factors positively affect sales (Hu et al, 2008; Hu et al, 2009; Hu et al, 2010; Lin et al, 2005; Lin et al, 2007).

These are the metadata of the book, and in traditional publishing, the original source for the metadata of a book is the publisher. The reader may come to the metadata

from a publisher in a few different ways, as illustrated in figure 1 by following the red links indicating the flow of book metadata. If the book is in print format and stocked at a local bookstore or available in a library, the reader may first come to it directly through physical contact with the book. Physical browse alone accounts for a substantial portion of total book sales, and publishers often compensate booksellers for premium product placement and display to promote browse behavior (Thompson, 2010).

Alternatively, publishers avail themselves of mass media to make information about books generally available to the public. They do this through marketing and advertising, through trade journals published for this purpose, and through release of information about books and their availability to mass media outlets for review, product placement and discussion.

At this point, there is the additional gatekeeping role held by the press and other mass media outlets including radio and television (and increasingly today, internet-based mass media channels), who decide which of the publishers' books to feature and which to ignore. Some readers may come to metadata either directly from marketing and advertising in the mass media, or they may come to it through the mass media secondary gatekeepers such as media reviews.

After its initial release by the publisher to the mass media via marketing and advertising, information about a book and some of its particulars can begin to flow by word of mouth into social communication channels through processes described in the previous section. Figure 2 (next page) shows the additional points from which information can penetrate social channels in today's book value chain compared to Figure

1. Prior to the widespread adoption of internet and Web technologies, there were relatively few ways that publishers, not to mention authors, could bypass mass media intermediaries and directly penetrate social networks. Publishers might engage personally with known book club leaders in face to face outreach, for example, or perhaps provide review copies to influential reviewers and commentators, which serve the purpose of influencing opinion leaders directly. And publishers have also experimented to a limited degree with consumer-direct models, such as book of the month club-style direct marketing.

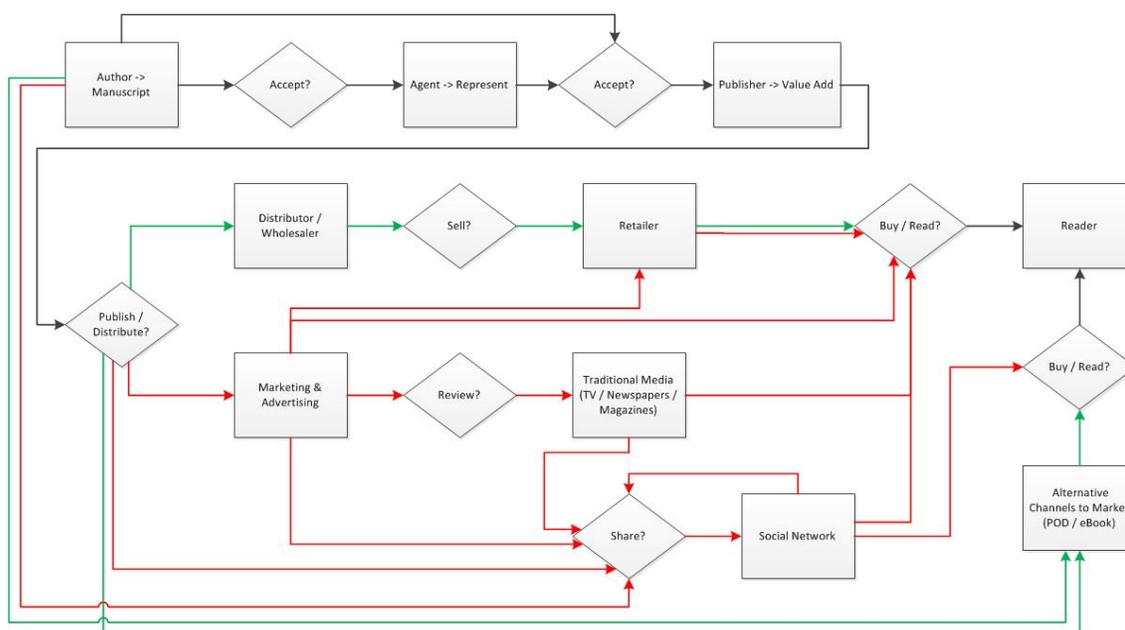


Figure 2 - Extending the Publishing Chain

Ubiquitous computer mediated communication channels, however, facilitate the potential direct communication with anyone with access to the internet and connected computer mediated communication channels – which today includes nearly everyone.

Some popular media pundits have suggested that the internet effectively bypasses the mass media gatekeeper intermediaries and signals the end of gatekeeping (Weir, 2011). Yet while it is true that technological innovation has lowered or eliminated barriers to self-publication and spawned new business models, other argue that traditional publishing and mass media gatekeepers serve a valuable role in selecting quality material, filtering out poor quality, and organizing selection characteristics for the reader (Thompson, 2010). Gatekeeping still serves a critical function in the process of peer review in scholarly communication and in professional journalism and other disciplines (Glogoff, 1988; Shoemaker, 1991; Shoemaker et al, 2001; Shoemaker & Vos, 2009). The book reviewer as gatekeeper has been explicitly noted in the context of scientific journals and scholarly publishing (Crane, 1967; Glogoff, 1988; Okerson and Magge, 1994). So it seems premature to declare gatekeeping dead.

In fact, it is true that the same kinds of technologies that lower barriers and reduce costs for self-published authors and niche publishers also lower the barriers for individuals to select, filter, add value to and diffuse information in computer mediated information channels. Rosenbaum, in his 2011 book *Curation Nation*, argues that consumers desire information filtered by knowledgeable individuals in order to help sort through the glut of information available. Traditional gatekeepers can continue to fulfill this role, but so can other individual amateur or professional intermediaries. In Rosenbaum's view, the expert is not necessarily the credentialed professional. Rather, an emerging cadre of informal and amateur experts can take advantage of the low barrier of entry to new communications channels and curate information for others by pulling the

best, most relevant information from the information glut and packaging it for the information seeker:

“Curation is about something different than disintermediation. In fact, it’s about remediation. It’s about adding quality back into the equation and putting a human filter between you and the overwhelming world of content abundance that is swirling around us every day.” (Rosenbaum, 2011, Kindle ed. Loc. 262)

The term “curation” in the sense Rosenbaum uses it sprang to popularity in 2010 according to Google trends (<http://www.google.com/trends/explore#q=curation>) and continues to be a popular meme, especially among companies attempting to implement Rosenbaum’s strategies for improving business through his model of curation via social networks. However, Rosenbaum’s strategies aren’t particularly novel, consisting primarily of selected examples of organizing and sharing information on blogs and social sites. The term “curation” is also problematic in the context of the academic community of curators and archivists who have developed disciplinary standards and practices. One key distinction is that curators traditionally manage an actual collection, either physically tangible or instantiated in digital form. Curation, as described by Rosenbaum on the other hand, is primarily concerned with the management of information *about* information objects or artifacts, or processes and ideas, that is, metadata with links or other connections leading to a tangible information object or artifact.

This is a key point of understanding: curation in Rosenbaum’s sense is, in fact, a type of gatekeeping practice in which individuals create information gates such as a blog, Facebook post, Tweet, review or some other kind of message in a social communication

channel and in doing so control, select, modify and filter information in the form of those messages that flow through the gate to individuals seeking them.

**Proposition:** Social Gatekeeping can be initially defined as the process of finding, selecting, filtering and shaping information about a product, service or idea and making it available (or not) as a message accessible in a social communication channel.

*Identification and Location of the Message as a Unit of Analysis*

In order to deal both quantitatively and descriptively with the message as unit of analysis, it is necessary to determine how to uniquely identify them. The information gate, be it a blog, Facebook post, tweet or review, is a message, and in computing terms, each uniquely identifiable resource can be identified by a Uniform Resource Identifier, or URI. So, each information gate, comprised of a message, is uniquely identified by a URI. The purpose of a URI is to uniquely identify a name or Web resource. There are different categories, or subclasses of URI's and the one most important to the study at hand is the URL (Uniform Resource Locator), which both identifies and locates resources on the World Wide Web. For example, <http://my.domain.com/mydoc.doc> identifies a particular resource with the name mydoc.doc, located on a server within the domain my.domain.com. <http://my.domain.com/myblog/todays-message> locates a particular blog message located and retrievable from the blog existing at my.domain.com/myblog. Other forms of unique message identification include other forms of URIs, or identifiers.

Individuals (and automated processes designed for the purpose) who act to find, select filter and shape information are social gatekeepers. Social gatekeepers may either

be intentional, that is, may be creating messages for the purpose of information diffusion (such as a dedicated book blogger or reviewer), or incidental, that is may be creating messages in the course of regular social communication that happen to include information that is passed on to others. When messages are created and disbursed using computer mediated communication, they can be uniquely identified and retrieved with the URL that is associated with the message.

Suppose, for example, blogger A becomes aware of a new book and writes a review of it. The review is posted to the blogger's site in the form of a blog entry, which is a Web page within the author's blog space uniquely identified by a URL. The blog entry is, as a practical matter, a digital gate through which information passes from the blogger, according to the blogger's personal filters, opinions, choices and preferences, to individuals who happen upon it and reflect on the contents. This is social gatekeeping in its most basic form, and blogger A acts in the role of social gatekeeper.

A reader may come upon the blog entry through browse activity and serendipitous discovery. The reader may subscribe to the blog for example, or follow the blogger's twitter account which links the blog, or the reader may happen across a comment by the blogger on another blog that traces back to the original blog entry.

However the reader discovers it, she then may propagate the information on her own blog or elsewhere, perhaps on a social network such as Facebook or Goodreads.com. In doing so, she may further filter the message, re-cast it, or add value to it. In effect, the reader becomes a social gatekeeper herself and by creating a new information gate on the Web increases the overall Web presence of the book and some of its metadata.

**Presence**, such as Web presence further developed in the pages that follow, is defined here to be the set of messages each identified and retrieved by unique URLs directly relevant to an individual, product, service or idea.

That is, her post, whether on her own blog, or as a message in a social network or in some other format, contributes to the overall Web presence of the book as measured by the total number of messages directly relevant to the book available for retrieval on the Web.

The effect of social gatekeeping on Web presence is to filter information in to the Web ecosystem, where it may be discovered, passed on and acted upon. A key distinguishing feature of social gatekeeping therefore, is that while traditional editorial gatekeeping tends to filter out the many available information messages in order to make a small number of them available to the information seeking public, social gatekeeping has the opposite effect: all of the social information gates in total contribute to the presence of an individual, product, service or idea within a network and makes the information discoverable.

### *The Serendipitous Tie*

Social network analysis, previously described, conceptualizes individuals as network nodes and analyzes their relationships according to a variety of traits including degree of homophily and tie strength, often measured by communication patterns and the particulars of information shared. However, social network analysis doesn't deal explicitly with the particulars of relationships that occur outside the purview of the

network structure, for example, communication between individuals with no known social tie. In the example given, both the blogger A and the reader cum blogger B may be connected in a traceable way through social networks; that is, they may have either strong or weak ties to each other that are apparent through connections established within social networks such as friending on Facebook, following on Twitter, or by other explicit means such as being subscribers to each other's blogs. If so, standard network analysis methodologies and theories can be used to measure and explain the patterns of communication and information diffusion within the confines of the social network under study.

However, it is also entirely possible that there is no a priori identifiable relationship at all between the two bloggers, and if the second blogger, in sharing the book information, does not explicitly link to or reference the first blogger, then there is no connection that network analysis can trace, even though the flow of information from A to B takes place. In this way, information appears to jump from one network or sub-network to another through ties that are more tenuous even than weak ties. Current research that uses social network analysis methodologies often treats these events as anomalous and may even attempt to control for them as confounds (e.g. Bakshy et al, 2012). Because some research shows that information can become trapped within social networks even given the strength of weak ties (Onnela et al, 2007), understanding this aspect of information diffusion is important to and may extend many of the current ideas concerning tie strength and its impact on the spread of information among and between social networks.

Granovetter's (1973) original research on the strength of weak ties considers the dyadic relationship tie as a continuous variable ranging from strong to weak, or absent. A footnote sheds light on the latter, noting that the *absent* tie label could include both the lack of any relationship and also ties with insubstantial significance, such as a "nodding" relationship (pg. 1361, footnote 4). Granovetter notes that negligible tie relationships might only be usefully distinguished from absent ties in rare circumstances such as disasters, implying that there is little or no communication flow between these individuals.<sup>2</sup>

However the nature of computer mediated communication and networks such as the Web makes the negligible or absent relationship tie potentially significant. In fact, among and between members of social networks and individuals on the Web generally, there can be useful and dense exchanges of information with no discernible degree of overt tie strength. This is the case with blogger A and reader cum blogger B, who may only interact through a serendipitous encounter with no network trace of the exchange of information, even though the information exchanged might be significant and meaningful.

In this example, A and B participate in a computer mediated social interaction that appears to sit outside the bounds of the strong to weak tie continuum.

---

<sup>2</sup> The face-to-face spread of information among negligible or absent ties does not appear to have been explicitly or methodically investigated, but almost certainly occurs more than Granovetter envisioned. For example, two individuals unknown to each other might strike up a conversation while waiting in the checkout lane at a supermarket and exchange information that is then returned to the individuals' respective social networks.

**Proposition:** The *Serendipitous Tie* is an incidental, chance or accidental interpersonal relationship event that may occur between people not otherwise socially connected, by means of which information may be passed and communicated from one individual, and potentially one social network, to another individual and social network.

This kind of interaction in which no prior or traceable tie exists represents an important information diffusion event that accounts for some portion of the diffusion of information among and between social networks that are otherwise treated merely as external.

In fact, A and B may never even interact at all if the information exchange is mediated by a process or rule invoked by an algorithm that results in a message that propagates through a social channel. Suppose A and B both read book X and give it a positive rating on a vendor site, which has been designed to enable social interaction. Then suppose A subsequently purchases book Y and gives it a positive rating. The next time B visits the site, she may be directed to a page or page element that suggests people who liked book X also bought Book Y. This is also a form of social gatekeeping, one that leverages social data and is mediated by a process rather than a person.

A second form of serendipitous information exchange can occur through friend of a friend interaction combined with social information processing that doesn't involve network traceable relationships. Suppose A sees B's friend C mention a book on B's Facebook wall. A proceeds directly to author X's public Facebook page, reads it and "likes" it in order to receive more information. An automated Facebook book recommendation application might note that A has liked a certain set of pages and from

there may recommend a particular book written by Y because the recommendation engine has a large enough database of “likes” and messages from a critical mass of users to create a business analytic that is able to link A with book Y through a series of serendipitous exchanges of information between individuals with no discernible traditional social tie. A may then explicitly post on her own Facebook page or elsewhere, and further propagate information as an intentional or incidental social gatekeeper.

Such automated recommendation engines that use social data and rule-based processing to find, filter and select individualized information are commonplace, and may be referred to as social gatekeeping applications. Serendipitous exchanges of information among individuals without traditional network strong/weak tie relationships may comprise some significant portion of information exchanges through a variety of different scenarios. Facebook alone has dozens of specialized book recommendation applications (“apps”) that leverage social posts (Boog, 2011). BookRX (<http://books.knightlabprojects.com/>) is an app that combs the content from twitter posts for an individual’s areas of interest in order to make book recommendations. Random House Publishing has launched what it is calling a “discoverability” application BookScout (<https://www.facebook.com/appcenter/bookscout>) that uses an individual’s facebook posts and “likes” to help readers discover new books. Some recommendation algorithms are hybrids, combining both professional gatekeeping practices and social data. Bookish (<http://www.bookish.com>) is an experimental book recommendation site that combines social data with other factors such as professional reviews and book awards. Bookish is financed by three of publishing’s largest conglomerates, Simon &

Schuster, Penguin Group USA and Hachette Book Group. Kaufman (2013) quotes Bookish CEO Ardy Khazaei as saying that that friends and relatives "won't be able to know about as many relevant books as our tool can." These kinds of applications all leverage social information to filter, shape, add value to and propagate information about resources and thus serve a social gatekeeping function.

### *Measuring Social Gatekeeping*

The serendipitous tie effect on the diffusion of information between and among discrete social networks on supernets such as the Web suggests that tracing information flows within individual social networks – even very large social networks such as Facebook – doesn't provide a complete look at the flows of information across all of the social networks and among unaffiliated network users (those individuals with network access but lacking membership in a formally organized social network space). However, although the information transfer among serendipitous ties may not be recorded in the network itself and thus cannot be tracked explicitly through network traffic analysis, the appearance of a new information gate, in the form of a new message, is evidence that it occurred. The general purpose search engine, while not especially effective at facilitating serendipitous browse, may nevertheless serve as a tool for measuring the penetration of information across the Web and its rate of change.

Consider that another way a reader may discover either the book or blogger A's message is through a general purpose search engine such as Google, Bing or Yahoo Search. The content of the blog message along with the URL is captured by search

engines and made accessible via search according to facets such as key words, phrases, content type and full text indexing. So if a reader uses a search engine effectively and with the right set of search terms, the blog, or perhaps directly, the book, may be discovered.

Unless a search is very specific, however, a search engine query result set may number in the tens or hundreds of thousands or even millions of pages (URLs), and any particular blog or book may not bubble up to the surface. That is because general purpose search engines are programmed to filter and display the most relevant resources (Brin & Page, 1998). This makes them useful for identifying the most popular and appropriate resources for a given query but limits their usefulness in finding the less common, less popular, alternative or niche resources. Further, most search engines limit the depth a searcher might actually navigate to a few hundreds or thousands of results out of all the Web pages indexed by the service. That is, not all of the pages indexed by the search engine are exposed explicitly and retrievable through the search result. So even if some particular resource matches the searcher's query, it may not appear in the limited number of results a searcher can actually peruse.<sup>3</sup>

---

<sup>3</sup> On Google, for example, it is possible to jump to lower ranked results by manipulating the URL Google uses to parse the return. However, attempting to jump to the 2000th resource of a query result generates this message (as of Winter 2012): "Sorry, Google does not serve more than 1000 results for any query. (You asked for results starting from 2000.)" Similar attempts on Bing resulted in varying numbers of searchable returns. But in fact, even if search engines provided complete results, it would not be practical to retrieve or sort through all of them if the returns are sufficiently large.

If the reader is looking for a new mystery book to read and enters the terms “mystery” and “book” into the Google search engine and requests a *verbatim* return, the total result count (as of Winter 2012) is listed by Google at over two trillion indexed pages, with the top visible results being sites about mystery books and specialized search tools to help the searcher locate mystery books, rather than specific books in the mystery category. As of Winter, 2012, the default *all results* Google return for some categories, such as searches with the term “book,” includes, in part, a scrollable list of popular books, based on page rank, thus effectively promoting the top rated or most popular books found in Google’s index. But the number of unique titles that can be searched by scrolling through the search results, limited to perhaps 1,000 pages, is a fraction of the universe of mystery book titles, so the great majority of book titles are not exposed to browse using this search technique.

On the other hand, general purpose search engines are adept at returning relevant results for very specific queries. If an eBook author and title are known and if the eBook is available, search results will almost always return pages relevant to the particular eBook with high ranking (early in the return result), and if the query is specific enough, the precision and recall of the result set can both be high.<sup>4</sup> Search result sets that are high in precision and recall reflect the distribution and number of URLs relevant to the query.

---

<sup>4</sup> Precision and recall are used in the standard information retrieval sense: For a given result set generated by a query on a collection of records, precision is the ratio of relevant to non-relevant returns in the result set and recall is the ratio of relevant items returned compared to the total number of relevant items in the collection on which the query was executed. High precision / high recall results consist of all or mostly all relevant results and include all or nearly all of the relevant results available from the collection of indexed records.

In effect, what is being returned with such a query given high precision and recall is a listing and count of the publicly accessible relevant information gates among all the sites indexed by the search engine.

Queries that return URLs referencing resources directly relevant to specific products, services or ideas enumerate the extent of publicly accessible information gates and thus provide a measure of the output of social gatekeeping. Given a sufficiently specific and appropriate query, therefore, the general purpose search engine is therefore a tool that in principle can be used to estimate, at a given point in time, the presence on the Web of products, services or ideas, including resources such as books and eBooks, where presence is defined to be the set of messages identified by unique URLs directly relevant to the product, service or idea. The search engine can also be used to track over time the change and rate of change – positive or negative – of the set of messages as the information diffuses through and among the networks publicly indexed with the information.

Not all gatekeeping messages are identified with a Web URL indexed by a search engine, so other methods must be used to track them. Messages may be uniquely identified either with an internal or public URI that is not an indexable Web page. Consumer reviews, for example, are often aggregated at a single URL by a social network host and displayed according to a reader's search criteria. In these cases, the social site almost always provides a summary of reviews including count by rating, and most social networks provide software interfaces through which individual reviews can

be retrieved if more than a simple count is desired (or screen-scraping methods of gathering information directly from screen displays can be employed).

It seems unlikely that a method or group of methods could be developed that would definitively, inclusively and orthogonally total all of the messages of interest across all networks. Search engines such as Google and Bing do not index every page and the surface Web available to indexing robots is only a fraction of the total number of publicly available pages and databases. However, absent hidden bias, search engines can be used to compare items for overall Web and network presence and thus demonstrate the effects of our understanding of the diffusion of information through networks and the effects of social gatekeeping.

#### *Social Gatekeeping as Strategy*

Figures 1 and 2, illustrating the traditional and emerging publishing chain, show the two diverging classes of information that pass through gates in order to reach readers, namely, the book itself and the metadata, that is, the information that describes characteristics of the book. As previously discussed, technology has virtually eliminated the barriers to publication of a book in either print or digital format, and the market is such that anyone who becomes aware of a specific and available book by author and title can easily acquire it.

Traditional methods of pushing the metadata out to the reader still present gatekeeping barriers, however. Mass media marketing both through traditional channels – such as broadcast television, radio and print advertising in newspapers and magazines –

as well as advertising in emerging computer-mediated advertising channels remain expensive, prohibitively so for most self-published authors and small/niche publishers without national advertising budgets. Saturation marketing campaigns are out of financial reach for all but the largest media companies promoting what they believe will become potential blockbuster titles. And, even the larger publishers are quick to cut off advertising support for titles that fail to meet initial sales targets, sometimes in as little as a few weeks following initial release (Thompson, 2010).

Breaking through to professional reviewers also continues to present significant gatekeeping challenges. The numbers of titles professionally reviewed annually is difficult to determine with precision, but the major review sources such as the few remaining book review newspaper supplements and the declining handful of dedicated trade and book review publications number only in the few thousands of title reviews each per year, with considerable overlap.

Two-step flow of communication-based theories suggest that mass media influences consumers through two broadly defined channels, first directly to the consumer through exposure to mass media messages, and second – and more prevalently – through persons of influence, who intermediate between the mass media message and the consumer. Information then diffuses among and between social networks through the methods and channels discussed above.

Mere publication of a book through a publishing service such as Amazon, Barnes & Noble or Smashwords results in a limited amount of Web presence in the form of

pages on the retailer site, which may be enough to reach some social gatekeepers.<sup>5</sup> But as figure 2 suggests, direct access to social networks through computer mediated communication channels may provide an alternative path to reaching a critical mass of social gatekeepers, with low barriers to participation. In essence, what authors and publishers may be able to do is to trigger an ad hoc network of social gatekeepers. These gatekeepers serve as an alternative to mass media by creating messaging directly at the edges of social networks where information diffusion can begin.

---

<sup>5</sup> Note, however, that some online retailers such as Amazon limit browse results in the same way as previously described for general search engines. As of fall 2012, Amazon limits the total number of pages that can be browsed based on a filter such as genre or subject to 100, with 12 results per page. Thus, books that fail to rise to sufficient levels of relevance are, as a practical matter, not accessible to browse strategies and will not be discovered absent other kinds of list promotion.

## Research Questions

The conventional wisdom often repeated in publishing circles and in the popular press is that maintaining personal Web presence and participating in social networking is an effective way for authors and publishers to initiate a flow of information about a book and its availability through their social connections (Skerik, 2011). Social gatekeeping describes the process by which individuals, if successfully reached by a message that aligns with their interests, needs and predispositions may in turn filter, select, and re-share information. Social gatekeeping suggests that a key strategy for authors and publishers, and especially for those without the means to engage mass media, is to get the message either directly to consumers if possible or to social gatekeepers who in turn will influence consumers.

These considerations flow logically from the theoretical perspectives and theory extensions presented in the previous sections. Creating author Web presence by managing Web sites and blogs and participating in social media, when used as tools to initiate the flow of information about an author's books, is one set of strategies among others that may increase both discoverability and sales. Research could show, therefore, that authors and publishers who place greater numbers of messages at network edges through author Web presence and social media participation will be more successful at propagating a cascade of sharing than those who do not.

The research agenda here is intended to ascertain whether empirical evidence might support the social gatekeeping model by confirming an association between the

dependent variables of book Web presence and sales as a function of independent variables of author Web presence, as suggested by the theoretical framework.

Although the concepts and terms discussed so far generally apply to both print and digital title formats, the research described in the pages that follow will focus on a random selection of eBooks newly released during a single week and tracked over time, discussed in more detail in the chapters that follow. A random selection of eBook titles includes titles for which a print and digital version are released as well as titles for which only a digital version (eBook) has been released. eBooks are acquired principally through a computer mediated communications network (e.g. an online Web store, download site or directly through an eBook reading device to a publisher's portal) whereas marketing and sales for print versions of titles may involve the tangible physical form, e.g. bookstore end-caps and display tables. Physical browse for titles in print may also impact sales of the digital versions (Digital Book Wire, 2012). By selecting a sample for research that includes both, analysis of the data may shed light on whether Web presence and the impact of author Web presence differs for physical vs. digital media.

In addition to differentiating titles on the basis of print vs. eBook format, other factors may affect the social diffusion of information about a title over time, and a random sample may be used to focus more sharply on some of the differences in order to make like to like comparisons. Some genres and subjects sell better than others, for example, and will appeal to more or fewer readers and therefore to more or fewer social gatekeepers. Other differences that might account for variance in diffusion of information through social gatekeeping might include book length and price, whether it comes from a

previously published or new author, whether the title is self-published or comes from a major or niche publisher, and other factors. The first portion of the research agenda is therefore descriptive and is based on a random sample of the current production of eBook titles.

RQ1: What comprises the current output of eBook production? What are its characteristics and alternative formats?

RQ2: How do the characteristics of the current output of eBooks break out by subcategory including genre and subject, length and price, self-published vs. mainstream published, and other factors?

Appropriate and systematic classification of eBooks sampled for research helps assure that appropriate comparisons are made. Further, this portion of the research creates baseline against which future research can be compared. It creates a starting point for longitudinal research of the emerging eBook market over time and grounds basic research in eBook authorship and readership with a snapshot of the eBook landscape at an emergent point in time.

Book Web presence and sales can be tracked and analyzed over time and are the dependent variables that can be measured as a function of author Web and social media presence. For purposes of this phase of the research, Web and social media presence include maintaining a Web site (.com or .net), maintaining a blog, either in conjunction with or separate from a traditional Website, maintaining a publicly accessible Facebook account including a count of friends or likes, maintaining a Twitter account including

counts of posts, following and followers, and registration as an author on Amazon and Goodreads, which puts those sites' registered authors and registered readers in virtual proximity.

RQ3. To what extent are eBook authors and publishers establishing Web presence, and is author Web presence differentiated by sub-category of book?

RQ4. To what extent does author Web presence account for search engine page hits and sales? Does its effectiveness vary by title sub-category, such as genre or self-published vs. mainstream published?

RQ5. Is there a relationship between sales, measured by sales rank, and Web presence, measured by search engine links returned?

These research questions and the social gatekeeping framework suggest these specific hypotheses:

H1: Author Web presence associates positively with eBook Web presence, as measured using search engine result counts with high precision queries.

H2: Author Web presence associates positively with eBook sales, as measured using Amazon Sales Rank

As noted in the previous sections, some kinds of social messages are not individually accounted in search engine estimates of message counts. These include, for example, consumer review counts on sites such as retailers Amazon and Barnes and Noble, and social book sites such as Goodreads. Consumer reviews are an important

component of social gatekeeping message creation and this raises certain questions and a hypothesis:

RQ6. What is the relationship, if any, between book Web presence and consumer review count?

RQ7. Do rates of diffusion of information on the Web and rates of numbers of reviews correspond over time?

H3: eBook Web presence associates positively with consumer review counts.

The results of this portion of the research quantify the relationship between the independent variables, an author's Web presence and use of social media, and the dependent variables of sales rank and search engine links, markers of readership and discoverability. The analytical techniques used are multiple regression and time series analysis.

The quantitative portions of the research include the largest of the well-known social networks including general purpose social networks Facebook and Twitter and the social book network Goodreads. However there are a multitude of other general purpose and specialty social networks where books may play either an intentional or incidental role. The third and final phase of the research consists of an extended open-ended review of author Web presence and use of social media in a selected subset of the titles selected for research beyond those identified as independent variables in the first two phases of research. The research questions guiding this portion of the research include:

RQ8. What additional insight can a more thorough examination of selected titles provide that informs interpretation of the results of the descriptive and inferential portions of the analysis?

RQ9. What does a more thorough examination of selected titles suggest for future research?

## **Nature of the Study**

This research project was primarily a quantitative study of books and their characteristics and the relationship of book Web presence and sales as a function of author Web presence. The data for the research came from a randomly selected sample of approximately 500 eBook titles drawn from the pool of approximately 9,600 eBook titles released on Amazon.com between March 30, 2012 and April 6, 2012.

The complete pool of most popular eBooks available on Amazon on April 6, 2012 as listed by Amazon was also collected for analysis. While not a control group per se, the pool of most popular eBooks collected at the same point in time serves the purpose of exemplar and best practices for comparison purposes.

The books from both the random and popular samples were categorized along several dimensions, including author Web presence as noted in the Research Question section above and as further explained in the section on Methodology (Chapter 3).

As suggested by Katz (1957) who emphasized the importance of studying diffusion over time in the context of an entire community, the Web presence of eBook titles from both the random and popular samples was tracked weekly during the period April 6, 2012 to July 20, 2012, a period of 15 weeks including 16 measurements. Web presence included the number of search engine hits on specific search engine queries, weekly sales rank as reported by Amazon and counts of reader submitted reviews on Amazon. Multiple regression was used to study the degree to which author participation in social media (author Web presence) was predictive of ebook Web presence.

Finally, the research included a qualitative review of a non-random selection of titles from both the random and popular samples in order to discover additional author Web presence strategies beyond those included as independent variables in the quantitative analysis. This portion of the research is intended to inform future research and aid in the interpretation of the quantitative results. The qualitative portion of the research is intended to be informative but not exhaustive.

## **Assumptions**

The research results and conclusions depend on the assumption that the variables used for the analysis have external validity and are suitable indicators for the underlying hypotheses. In particular, the two primary dependent variables used are proxies rather than direct measurements.

In the case of sales, accurate sales data for eBooks is both proprietary and incomplete. Nielsen's BookScan, the publishing industry's primary book sales tracking service, tracks only certain major retailers and e-tailers for sales data (Charman-Anderson, 2013), which they sell to publishers, authors and other industry partners. Access to their proprietary database would be prohibitively expensive for a study such as this and still would be incomplete. Amazon itself does not publicly disclose actual sales.

Therefore, the measured dependent variable for purposes of this research is Amazon sales rank, which is an ordering of books according to sales numbers where the actual sales figures can only be imputed. The primary criticism of this variable is that the formula for determining sales rank is proprietary. However, independent researchers have reverse-engineered the algorithm and established a relationship between the sales rank and sales. Thus, sales rank has been used previously in peer reviewed research as a proxy for comparing sales across several markets. Strictly speaking, Amazon sales rank is only a proxy for sales on Amazon, so care must be taken in generalizing the results (see Limitations below).

A similar situation exists with regard to search engine counts, which purport to represent the number of documents (URLs) indexed by the search engine relevant to the search term. Search engine result counts from the two major search services, Google and Bing, are estimates reported only to three significant figures, and the algorithms used to generate the estimates are proprietary, so there is no way to directly confirm results in any particular case. However, as with sales rank, independent research on search engine count validity has established that the results are generally reliable under specific conditions, and search engine counts have also successfully been used in peer reviewed research.

As used in this research, both sales rank and search engine counts are used to compare differences among eBooks by category based on author Web presence, so the precision of the two independent variables is less important than their overall reliability, on the assumption that neither dependent variable is biased in some hidden way with respect to the independent variables. That is, there is no reason to believe that among the group of self-published romance eBooks in either the random or popular sample, for example, sales rank and/or search engine count contain a hidden bias related to author participation in social media other than the proposed hypothesis itself, that search engine count and sales rank associate positively with author participation in social media.

The literature review chapter provides additional insight into the background and use of sales rank and search engine count as independent variables.

### **Limitations and Significance of the Study**

While the primary analytical tool used for this research, multiple regression, can be useful at predicting the relationship and relative strength of independent variables on dependent variables, it does not conclusively establish causality. It is also susceptible to independent variable selection bias and may be misleading as model complexity increases.

Critics of non-experimental methods, including multiple regression, suggest that randomized field trials are a better choice for hypothesis testing. However, a randomized double-blind trial designed to control for hidden factors and confounds would be difficult to devise and execute in the book publishing business. The results of the regression analysis indicate support for the research questions and hypotheses, and are themselves triangulated by evidence from the descriptive and qualitative portions of the research. They initially frame a line of inquiry for further research that could include more difficult (and expensive) field trial experiments.

The research examines social gatekeeping and the serendipitous tie as a framework for understanding and analyzing the flow and diffusion of information among and between social groups and networks. Social gatekeeping, as an extension of traditional gatekeeping theory, and the newly introduced concept of the serendipitous tie have potential applicability beyond that of the publishing industry, and the research therefore is of broad and general interest to other disciplinary communities for whom social diffusion of information is important. In that respect, the research may provide evidence that underpins a better understanding of social gatekeeping and may suggest

areas for future research including experimental approaches that could add to the evidence for or against it and its broader applicability and acceptance beyond the confines of the study at hand.

A second limitation of the study is that the results and interpretation are only strictly generalizable to books released and sold by Amazon under the Amazon rules and agreements in effect at the time the random and popular samples were drawn. At the time of data collection for this study, Amazon was responsible for the majority of sales in the eBook market at an estimated 60% (Streitfeld, 2012), but there are an unknown number of books released by and available for sale at outlets other than Amazon that might generate different results.

Limiting the study to Amazon data is primarily a result of the difficulty in generating a random sample selection of titles from other sources. Of the major sources of currently published eBooks, including vendor-publishers such as Barnes and Noble, Smashwords, and several others, only Amazon was found to provide a search-browse function that could return a complete population of books in a non-biased return order (in this case, date of publication). Further, Amazon is the only one of these vendor-publishers that provided robust computer-based access to the eBook's internal metadata. Some otherwise authoritative sources of book data, such as Bowker's are incomplete. For example, the Bowker database does not include titles published on Amazon without an ISBN number.

Amazon currently provides the only venue in which it is possible to conduct a study of sufficient size and complexity to address the variables and constructs in the

theory and hypotheses of this research. At this point in time, it is the best operational choice to study the current universe of e-books. This situation will undoubtedly change, perhaps sooner rather than later, as other players become stronger and as eBooks evolve. When that happens, the results from this Amazon study will provide a baseline from which to observe changes.

These issues are explored more fully in the chapters that follow.

## Definition of Terms

**Book** – The Functional Requirements for Bibliographic Records conceptual model defines a *work* as “a distinct intellectual or artistic creation, independent of any concrete realization or expression of its content.” The *expression* of a work is the realization of the work in the form of a notation such as alphanumeric, musical notation, sound and so on. The *manifestation* is the “specific intellectual or artistic form that a work takes each time it is realized.” (Denton, 2007). A book in this context refers to a manifestation that instantiates the *work* in a physical form, such as an edition of a printed codex, of which individual examples are *items*. The term *book* is also popularly if imprecisely used to indicate the *expression* of the work, that is, the expression of the work regardless of its physical instantiation (Reitz, 2004). In this sense, *book* may refer to and be instantiated as a printed codex, or it may be an audio recording of the text of the work, or it may be the work formatted as a computer file or in some other manner.

**Book: eBook** – In the most general sense, an eBook, short for electronic book, is a *work* (see above) published in a digital format designed to be read or viewed on some kind of electronic device, such as a personal computer or a specialized electronic device such as a dedicated eBook reading device or tablet computer. An eBook may or may not have a print analog; the term distinguishes print and other physical instantiations from digital as a choice of format. The term does not imply a particular digital format or device; it may be coded in a general purpose format such as PDF (portable document format) or in a container format developed especially for the display of literary works on specific digital devices such EPUB (open standard), KF8 (Amazon.com proprietary), and

others. The term eBook is sometimes used elsewhere to refer to the device on which digital text is displayed, but used here, the term refers to the digital text itself and its digital encoding and not the physical device used to render digital text.

**Book: eBook app/application** – eBook apps are software applications programmed to render a literary work on a particular digital device. The software provides not only for the rendering of text but may also include features not found in print books such as multi-media audio and visual playback, hyperlinks, interactive reader features, and other features available on the host device.

**Channel (communication)** – a medium through which a message is transmitted or conveyed in order to reach its intended audience. Examples of mass media channels include broadcast radio and television, print media including newspapers and magazines, and advertising channels in online venues. Channels may also include private or semi-private media capabilities such as messaging on Web-based computer-mediated social networks, alternative computer mediated channels such as text and chat, and in person face to face direct communication.

**Discovery/Discoverability** – Strictly speaking, discoverability is the quality or degree of being discoverable. In practice, and in the context of this research, it is the degree to which a book stands out and can be found by a reader desiring a book that will fulfill a reading need. The term is not associated with any particularly specific measurable quantity. Virtually all books are discoverable by known author/title, whereas only a few of the many millions of available titles are practically discoverable through keyword search in a general purpose search utility. Book Web presence (see below) is

one available proxy for discoverability that addresses the possibility of finding some particular book through serendipitous browse and social network diffusion of information.

**Gatekeeper** – An individual or rule-based process who acts to find, select, filter, shape and propagate or reject information.

**Heterogeneous Marketplace** – The current marketplace for books, including traditional bricks and mortar book stores, chain stores, independent stores, and a variety of emerging online retailers, all selling a mix of traditionally published books as well as non-traditional and non-traditionally published literary works (Bradley et al, 2011).

**Hyper-abundant** – The sheer number of available titles and new title production which has increased by an order of magnitude or more from title production levels at the beginning of the 21<sup>st</sup> century (Bradley et al, 2011).

**Information gate** – A) In the publishing chain, the decision points in the flow of information from author to reader where information or content is passed through to the next decision point or rejected. B) In computer mediated communication, a message uniquely defined with a URL through which information enters the network.

**Message** – Generally speaking, a message is information passed from a source to a receiver through a communication channel. As used here in context, message refers either to the information itself or to the container object that carries the message. The containers for messages transmitted via the web, and through other forms of computer

mediated communication, are web pages or similar constructs that can be uniquely identified and retrieved with a URL.

**Metadata** – Generally speaking, metadata is defined as data about data. Used in the context here, metadata is any information that describes a book, as contrasted with the information that is the literary work itself. Metadata can include descriptions of contents, descriptions of physical characteristics, opinions, reviews and any other information that might help a reader decide to read (or not read) a book.

**Precision and Recall** - For a given result set generated by a query on a collection of records, precision is the ratio of relevant returns to non-relevant returns in the result set and recall is the ratio of relevant returns returned compared to the total number of relevant returns in the collection on which the query was executed.

**Print on Demand (POD)** – a business and distribution model using digital print technologies in which books are printed and delivered to a customer only when an order is received. Digital printing technologies makes printing even a single copy of a book economically feasible and eliminates minimum run requirements of traditional offset printing and reduces or eliminates costs such as warehousing.

**Publisher** – “The entity or individual who selects the material to be published, makes the decisions, and pays the bills.” (Bradley et al, 2011)

**Publishing: traditional/mainstream**– The model of book publishing prevalent from the 19<sup>th</sup> century to the present whereby a publisher acquires the rights to reproduce a work from an author who is compensated through advance payments and/or royalties on

sales. The publisher takes responsibility for printing, typesetting, artistic design, warehousing, marketing and all other aspects of the publishing value chain leading to acquisition by a reader (after Bradley et al, 2011).

**Publishing: Non-traditional** – In contradistinction to traditional publishing, a model whereby the publisher acquires or owns by virtue of authorship the rights to reproduce a work by a means other than advances and royalties. Such models include self-publishing, cooperative agreements between an author and other individual or firm (such as a literary agent) concerning financing of the publication of a work and the sharing of profits, author cooperatives, and other kinds of business and financial arrangements that fall outside the traditional confines of author advances and royalties (after Bradley et al, 2011).

**Publishing: Vanity** – A term applied to the model of self-publishing prevalent in the mid- to late 20<sup>th</sup> century whereby an author would contract with a fee for services publisher to prepare and print a manuscript in traditional book form at the author's expense. The term has a pejorative connotation. The term is currently used with far less frequency given the increasing acceptance of self-published works marketed through non-traditional channels, even though there is little distinction in practice between the vanity publishers of the late 20<sup>th</sup> century and the fee-based publishing services prevalent beginning in the 21<sup>st</sup> century.

**Publishing chain / publishing value chain** – the succession of intermediaries who act to select, filter and add value to a literary work as it moves from the author to the publisher to the market and finally to the reader.

**Search engine optimization** – the practice of manipulating variables used by search engine providers to rank site relevance in order to achieve a higher rank and increased visibility in search engine results.

**Serendipitous Tie** – an incidental, chance or accidental interpersonal relationship that may occur between people not otherwise socially connected, by means of which information may be passed and communicated from one individual, and potentially one social network, to another individual and social network.

**Social Network** – As used here, A) the set of friends, colleagues, acquaintances and other people with whom an individual communicates and interacts; or B) a community of individuals who communicate and interact through a computer mediated communication channel such as a Web site or other technology; or C) the Web site or other technology service that facilitates computer mediated communication.

**Web Presence** – As used here, A) the set of messages identified by unique Web URLs directly relevant to a product, service or idea; or B) when referring to a person (e.g., an author), the set of Web-accessible messages, posts and sites maintained by the author for the purpose of self-promotion and/or the promotion of products, services or ideas.

**Work** (n.) – a distinct intellectual or artistic creation, independent of any concrete realization or expression of its content. See Book.

## CHAPTER 2 - REVIEW OF THE LITERATURE

### History and Definition of eBook

The first examples of electronic books, or eBooks, as we understand them today, come from the Gutenberg Project founded in 1971 for the purpose of creating a digital library through conversion of public domain literature to electronic text readable on computing devices (Lebert, 2009). These first examples consisted of pure ASCII-encoded text often entered into a computer file manually by volunteers, and much of the Gutenberg library is still available in pure text format, although other standards for presentation and rendering have since been developed.

Conceptualizations of electronic libraries of books in formats other than the traditional paper codex precede the Gutenberg project by several years. For example, the memex device was proposed as a thought experiment in the final days of the Second World War (Bush, 1945). The memex was conceived to be a virtual library, no larger than a desk, into which an individual could store all of their books, papers, files and communications, mechanized so as to provide rapid retrieval. The proposed technology was based on improving microfilm. Another device, this time based on computer technologies often linked to early conceptualizations of eBooks is the Dynabook (Kay, 1972), envisioned as a personal computer for children loaded with libraries of books and references and looking surprisingly like a modern day Kindle e-reader based on early sketches. Some confusion exists even today between the eBook as text stored in digital files versus the eBook as the device, container or collection of works on a device. For

example, OCLC has some catalog records of Kindle devices distinct from the digital files stored on them (Pers. Obs).

Even accepting for the sake of definitional clarity that eBook is the literary work and not the physical device or container, there have been competing and sometimes opposing views on what the eBook is. Bennet (2006, 2011), arguing from the publisher's perspective, has proposed that there is no single definition for eBook. Vassiliou and Rowley (2009) have undertaken a review of the literature and point out that the confusion over what an eBook is may inhibit the adoption of eBooks within academic publishing.

Part of the problem is that there is not a clear one to one correspondence between aspects of the print book and the eBook. For example, magazines and serials are not generally thought to be books as commonly conceived (although a collection or volume might be bound in book format), but they are popularly released in digital form labeled as an eBook. eBooks may also be released without a print counterpart at all. A single short story of a page or two, or a single poem, or even recipe, is released today as an eBook. And as eBooks and eBook devices evolve into computer applications, eBooks may embody components and devices not possible in print book format. Vassiliou and Rowley therefore and in consideration of some of these issues arrive at a two part definition of eBook:

“1. An eBook is a digital object with textual and/or other content, which arises as a result of integrating the familiar concept of a book with features that can be provided in an electronic environment.

“2. eBooks typically have in-use features such as search and cross reference functions, hypertext links, bookmarks, annotations, highlights, multimedia objects and interactive tools.” (p. 364)

The second part of the definition clearly must be subject to ongoing revision as publishers and authors continue to experiment with what actually works in the market place. Indeed, although Part One of the definition refers to the content aspect of the eBook, Part Two incorporates aspects of the hardware device and software used to render the text.

Technology changes rapidly, with adoption at the whim and predilections of the consumer. Tian and Martin (2011) identify technology and market demand as major factors driving the recent rapid rise in eBook consumption, which in turn drives customer behavior and competition. And, government policy and regulations, such as the monopoly litigation between the justice department, major publishers and Apple computers over the agency model (Streitfeld, 2012) provide ongoing pressure on the market and market forces.

For purposes of this research, a relatively simple definition of eBook has been developed, as much out of practicality and necessity as anything else:

**EBook:** A work published in a digital format designed to be read on some kind of electronic device, such as a personal computer or a specialized electronic device such as a dedicated eBook reading device or tablet computer.

Even this limited definition may be challenged as technology and eBooks evolve in the coming years. Further, a small number of eBooks meeting this definition selected as part of the random and popular samples drawn for this research proved unsuitable for purposes of the current investigation and were excluded from the analysis. Examples included computer games released as eBooks, single issues of magazines, and non-traditional literary works without clear authorship such as collections of Web pages or Wikipedia articles.

## **Publishing and the Publishing Chain**

The publishing chain is examined in Thompson (2005) and further developed in Thompson (2010). Thompson describes the publishing chain as both a supply chain, which traces the production, distribution, wholesaling and retailing of the physical book and a value chain, the series of steps in the publishing process where various entities make some contribution to the final product that adds value. Examples of these include activities such as typesetting, art, design, quality control, printing and so on. Thompson notes the gatekeeper role played by the publisher and comments that while publishers do indeed filter out many books, they also filter in value and add to the overall quality of books that do reach the market.

Greco (2005, 2007, 2011) has summarized publishing data spanning decades and comments that the publishing value chain is inefficient. He cites the practice of allowing returns from booksellers of unsold books as one example. Addressing inefficiencies in the traditional book publishing industry has led publishers to experiment with non-traditional arrangements such as cost and revenue sharing with authors, and eBooks for which there are no returns (Greco, 2005; Thompson, 2010).

Small, niche and independent publishing statistics have been collected by Wharton and Greco (2004) whose numbers were estimated to be 80,000 in 2002, well before the current explosion in self-publishing and eBooks. While major publishing conglomerates comprise the lion's share of sales, the smaller publishers are responsible for a significant share of total title production and Wharton and Greco's estimates of sales indicates a robust long tail for small, independent and self-publishers.

Up-to-date annual sales summaries are available through R. R. Bowker, a Proquest Subsidiary. Bowker is the US agent for the issuance of ISBN numbers and reports annually on sales figures for authors and publishers who acquire one or more ISBN numbers for their books and eBooks (New Book Titles and Editions, 2012). These tables document the rapid rise in non-traditional publishing from 2002 to 2011, but are not complete because many eBook and self-published authors do not acquire ISBN numbers for their works (Kilborn, 2010). The full scope of today's publishing market appears to be unknown and difficult to quantify.

The state of the current publishing landscape, and non-traditional publishing in particular, has been studied and documented by Bradley et al (2011). They determined that Bowker's estimates of ISBNs issued to so-called non-traditional publishers included both self-publishers using fee-based publishing companies such as Lulu and Authorhouse, and a few companies dominating the royalty-free content market. The royalty-free portion of the market includes the public domain reprint market, where out-of-copyright materials are reprinted, sometimes with new copyright dates, and publishers of so-called phantom titles, which may be nothing more than digital files of text culled using automated website scraping techniques and computer-generated metadata. These texts are not instantiated, either as eBook or printed copy, unless and until actually ordered. One such company, BiblioBazaar, was credited with publication of 1,461,918 individual titles in 2010.

On the other hand, many self-published authors self-publish completely independently including establishing a publishing imprint identity, managing printing and

eBook formatting, producing or acquiring artwork, and other services of the trade without using an author services publishing company. These works appear under their own imprint and show up in Bowker's data base as traditionally published material. It can be difficult to tell a self-published work from the output of a small traditional publisher without careful study of the particulars.

A number of self-published books are never registered via ISBN with the Bowker agency. Kindle Direct Publishing has made it easy and economical for self-publishers to create eBooks marketed on Amazon which carry only the Amazon Stock Identification Number (ASIN). These same titles may also be released on other services such as BarnesandNoble.com which assigns a stock-keeping private ISBN not registered with Bowkers, on eBook publishing services such as Smashwords.com without any identifying number at all, and as print-on-demand titles with or without an ISBN.

This is a recent shift coinciding with a sharp rise in eBook title production and Amazon's entry into the publishing arena around 2009. A study of self-published books by Bradley, Fulton and Helm (2012) was based on a random snapshot of works published through approximately 100 fee-based publication services. The works consisted mostly of print-on-demand book titles with some digital files noted, usually a PDF copy of the digital print file. Over 60% of these books were identifiable through a Bowker-assigned ISBN. Most of the titles lacking an ISBN were published by Lulu.com and made available exclusively on the Lulu Web site. This research predates – for the most part – Amazon's entry into the publishing business, although Amazon acquired existing fee-

based publishers such as Booksurge, Customflix and eBook company Mobipocket and later re-branded them as book and eBook publishing arm Createspace in 2009.

Few self-published titles make it into libraries (Bradley, Fulton and Helm, 2012; Dilevko and Dali, 2006). Libraries remain an important source of discovery, even for eBook titles (Rainie et al, 2012; Chandler, 2012), so the lack of access to libraries by self-publishers puts them at a disadvantage in terms of discoverability.

Bradley et al (2011) refer to the current publishing milieu as a blurring of boundaries, characterized by a hyper-abundance of titles in a diffuse heterogeneous marketplace where traditional mainstream publishers, small and niche publishers and self-publishers all experiment with business models, new practices and emerging technologies in order to connect readers to authors and their books. They further identify discovery as a major challenge facing authors and readers and suggest that if publishing is to become democratized, authors have to find ways to construct sufficient pathways that lead readers to their books. Social networks including Goodreads and Librarything are identified as examples of potential channels useful for discovery.

## Gatekeeping

The concept of gatekeeping as a framework for message selection, filtering and diffusion was introduced by Lewin (1943, 1947). Lewin wanted to understand what would be the most effective means of changing food habits and suggested mass media, individual approaches, or a direct appeals to a strategic portion of the population as potential influencers. In examining the problem, he introduced the notion of social channels and people of influence within those social channels as well as the notion of channels through which food moved to the table (grocery store, garden, etc). He noted that food moved through the channel by virtue of decision makers based on information at critical stages within the channel, such as the stage where, for example, food was purchased. He called these decision points *gates* and the decision makers who acted on information to move food through the channel or block it *gatekeepers*. Lewin writes:

“The relation between social channels, social perception, and decision is methodologically and practically of considerable significance. The theory of channels and gate keepers helps to define more precisely how certain ‘objective’ sociological problems of locomotion of goods and persons intersect with ‘subjective’ psychological and cultural problems. It points to sociologically characterized places, like gates in social channels, where attitudes count most for certain social processes and where individual or group decisions have a particularly great social affect.” (Lewin, 1947, p. 146-147)

These concepts – communication channels, information flow, social influence and message filtering and diffusion – remain central to gatekeeping as it has come to be adapted to a number of disciplines in varying contexts.

Lewin's work suggested only in general terms that information flowed through channels controlled by gatekeepers subject to rules or individual differences and implied but did not directly explore the applicability of gatekeeping to other circumstances and disciplines. White (1950), a student of Lewin, applied the concept to the flow of news through an editor who filters, selects and diffuses stories through mass media such as newspapers. White observed that only a fraction of stories that cross an editor's desk make it to print and that what the public finally reads is subject to a chain of gatekeepers ending with the editor in chief whose decisions are ultimately (p. 390) "...based on his own set of experiences, attitudes and expectations [of what] the communication of 'news' really is."

Geiber (1956) disagreed with White (1950) in a study of wire editors, who he found were limited more by process and institutional conventions and less inclined to exert personal preference. Geiber's study thus found more evidence for Lewin's rules-based mechanism of gatekeeping and institutional values as opposed to individual values. This understanding was confirmed experimentally in Shoemaker, Eichholz, Kim and Wrigley (2001).

A third gatekeeping theory influence in communication and journalism around this time comes from Westley and MacLean (1957) who effectively harmonized two-step flow of communication theories with gatekeeping. They pointed out that in fact there were three spheres of gatekeeping influence: mass media, interpersonal communication, and direct experience, with feedback from receivers to senders as well as from gatekeeper to recipient.

Because most stories that come through the gatekeeping chain of reporter to wire service to news desk to editor in chief never reach the public, editorial gatekeeping is a process primarily of information filtered out, with individual differences primarily accounting for selection. Shoemaker (1991) initially defined gatekeeping simply as (p. 1) “the process by which the billions of messages that are available in the world get cut down and transformed into the hundreds of messages that reach a given person on a given day.”

The editor as gatekeeper filtering out all but a few of the many stories crossing their desk remains an important framework of understanding. Glogoff (1988) surveyed scholarly journal referees and found that only about half used some kind of objective evaluation criteria in their reviews, supporting the individual differences perspective of gatekeepers. Some studies show that while individual preferences guide individual differences, mass media may tend to limit the variety of perspectives presented to the public if gatekeepers as a group hold certain perspectives and preferences. For example, Hoskins and O’Loughlin (2011) analyzed how jihadist speeches were processed, intermediated and disseminated through US and British mass media news agencies and concluded that gatekeepers tended to homogenize, misrepresent and omit perspectives that results in impairment of objective understanding.

Gatekeeping therefore has remained a popular and often used framework for understanding how news reaches the public in the disciplines of Communication and Journalism. Shoemaker and Vos (2009) summarize the current gatekeeping theory framework primarily as it applies to these fields. Shoemaker and Vos’ main contribution

in this monograph is the notion of differing levels of analysis for various gatekeeping constructs. The most granular of these is the individual level of analysis, that is, the nature of the individual and the individual differences that lead gatekeepers to arrive at individual decisions. Each of the succeeding levels of analysis encompass increasing levels of social organization, from communication routines between and among individuals, to organizations, to social institutions and then finally to social systems. Each of these is framed primarily in the context of 20<sup>th</sup> century communication paradigms, organizations and social structures. Shoemaker and Vos suggest that the technological changes of the 21<sup>st</sup> century and advances in computer mediated communication may necessitate a new conceptualization of gatekeeping and the mechanisms used for analysis of content. In particular, they call out the need to understand and incorporate into the gatekeeping framework:

- rapidly changing environmental conditions
- gatekeeping activities in evolving social systems
- the impact of globalism on communication patterns
- news content [e.g. the message] as a dependent variable
- how to conceptualize new levels of analysis
- the rising importance of the individual as gatekeeper
- the use of new or previously unapplied statistical modeling techniques

- a closer look at the information gates themselves including their characteristics and driving forces
- analysis of the characteristics of messages
- similarities and differences of the gatekeeping mechanisms of the various communication organizations.

While editorial gatekeeping, as studied in communication and journalism, reflects the predominate popular view of gatekeepers as filters, the library and information science perspective also supports the role of the gatekeeper as a selector, acting to filter in resources that inform a variety of perspectives, and in some cases helping to preserve and maintain culture. Shearer (1981) addresses the issue of librarian as gatekeeper as selector and stresses the importance of putting aside individual differences in the librarian's primary role as selector. Noting the appearance of bias and personal preference in collection practices (calling out for example the appearance of disproportionate works on hobbies according to the individual librarian's predilections), Shearer calls for librarians to be a fundamentally different kind of gatekeeper than White's editorial gatekeeper, and to reflect (p. 93) "the less popular, the more specialized, and the deeper messages [that] may be found nowhere but in the library." In this sense, the librarian acts to bring information to the reader's attention that otherwise wouldn't be available or considered.

Lim (1995) suggests that gatekeepers and the gatekeeping function are the key links to connecting with communities of diverse cultures. Lim defines gatekeepers as (p. 18) "persons in a particular community (particularly with reference to ethnic

communities) who assist individuals in their community to gain access to resources that are needed to solve problems.” Lim therefore identifies gatekeeping as an information intermediation process and the gatekeeper as an information intermediary who often translates between and among disparate communities of people. These kinds of gatekeepers, according to Lim, may be formally charged with those responsibilities, such as librarians, or may be informal members of their community due to their personal characteristics, contacts and influence. These informal gatekeepers, are in effect, gatekeepers of their community, and functionally represent the individuals of influence described by Lewin and other gatekeeping and information diffusion theorists.

Su and Contractor (2011) extend the notion of gatekeeping as cultural intermediation to business culture, noting that gatekeepers refer to those individuals who possess and maintain knowledge of certain information and that these individuals are sought for their knowledge and are thus critical to information transfer. Sturges (2001) makes a similar argument for the emerging role of the librarian in digital culture and reprises the technological gatekeeper definition of Allen (1977), who wrote (p. 150), “The phenomenon of the gatekeeper is not an isolated one. Rather it is one example of a much more general class of phenomenon. There will always be some people who, for various reasons, tend to become more acquainted with information sources outside their immediate community. A large proportion of these people in turn attract colleagues from within the community who turn to them for information and advice.”

Joyce (1998) reiterates the cultural gatekeeper as information intermediary theme and addresses both librarianship and publishing in his case studies of three Black

librarians who became publishers of works relating to Black culture and society. Recognizing the paucity of publishing opportunities for Black authors and poets, these three librarians used their library training and background to create new publishing opportunities that filled niche market demand not met by other publishers. Although generally modeling mainstream publishing business models, these enterprises nevertheless served to filter in new information to underserved communities.

Barzilai-Nahon (2008, 2009) has conducted a critical literature review of gatekeeping theory across disciplines including not only the communication/journalism and library science perspectives noted above but also law, MIS, management political science, public affairs and sociology. A total of 453 articles in key journals from 1995 to 2007 were located from a pool of nearly 25,000 total articles. Analysis showed that few articles on gatekeeping use gatekeeping as a method of analysis or as a quantified factor. The majority of articles used gatekeeping as a guiding framework, metaphor or concept. Her review confirms the two primary rationales for conceptualization of gatekeeping as a) editorial and b) preservation of culture, as discussed above, along with a small handful of other rationales, but leads her to conclude that gatekeeping lacks an overarching theoretical construct.

Barzilai-Nahon's review of gatekeeping underpins her development of an extension of gatekeeping theory she calls network gatekeeping theory in order to address the current lack of a foundational theory. Meant to apply to gatekeeping processes within networks, such as the Internet, Barzilai-Nahon introduces the concept of *the gated* as the individual subjected to a gatekeeping process and proposes that the gated have salience to

the gatekeeper, where salience is defined to be the degree to which gatekeepers give priority to competing gated claims, along four axes: their political power in relation to the gatekeeper, their information production ability, their relationship with the gatekeeper, and their alternatives in the context of gatekeeping (Barzilai-Nahon, 2008 p. 1498). The gated are then classified and deconstructed according to the 16 possible permutations of presence or absence of the four factors. The theory suggests that gatekeepers moderate gatekeeping practices according to the particular set of characteristics of those subjected to their gatekeeping practices.

Network gatekeeping theory has undergone little empirical testing. Bui (2010) brings the network gatekeeping framework to a study online news portals Google and Yahoo, which aggregate news from a variety of original and non-original sources, then filter, select and cull from the large number of available stories in order to present a limited number of them, typically ranked in order of importance, to readers of their sites. Bui determined that certain predictions of network gatekeeping theory regarding the role of the gated relative to the gatekeeper and the power the gated should have over news selection and placement were not supported by the analysis, but that traditional gatekeeping theory, in which varying levels of individual and rule-based decision-making guide the selection and filtering of messages through their portals, were generally supported.

Social gatekeeping is not proposed as an alternative to network gatekeeping theory but rather as an overarching theoretical construct of gatekeeping at the individual and social level, with the individual acting in the capacity of gatekeeper within the

context of a social network. Network gatekeeping is primarily a theory to explain the filtering behavior of gatekeepers as a function of the power of the gated to influence gatekeepers, in particular, those with considerable gatekeeping influence such as the mass media. Social gatekeeping, on the other hand, attempts to provide a mechanism and framework for understanding how, what and to what extent the gatekeeping construct explains the diffusion of information among and between social networks and groups of individuals possibly but not necessarily connected by social ties as traditionally understood and measured. While traditional and network gatekeeping theories generally explain how information is filtered out so that billions of messages are reduced to a few, social gatekeeping theory suggests that the net effect of all of the many individual gatekeepers together tends to filter information in and make more rather than less information available for consumption.

**Discovery: Serendipity and Browse<sup>6</sup>**

The discoverability problem has been addressed in the library and information science literature primarily in the context of two information seeking behaviors of the library patron, search and browse. Traditional library classification, cataloging and shelving practices have evolved to fulfill both needs. Search is primarily concerned with the systematic retrieval of specific information for which a cognitive need is reasonably well defined. In contrast, browse is cognitively less specific and lacks a well-defined information-focused need (Reitz, 2004). Some evidence suggests that relevance in the context of search is cognitively different than the notion of relevance for the browser and that for browse, relevant documents are not necessarily those that match a particularly defined topic (Bodoff, 2005). Exposure to a wide variety of topics, or even to non-topical associations may therefore serve the browser where it would only distract the searcher.

Libraries with physical print books are ideally configured to support both search and browse. The classification and cataloging functions of the library facilitate focused search, whereas shelving strategies and physical layout can facilitate less focused browse activities. For example, school children may be taught how to look up specific information in the catalog (search), or they can be introduced to discovery via browsing through exercises such as looking for a book on the bottom shelf, or finding an interesting book with a green cover (Coleman, 2007).

---

<sup>6</sup> Portions of this section reflect Fulton, 2010a, unpublished.

Library search, discovery and retrieval has moved increasingly online with the conversion of journals and reference materials from print to digital and with the move from card catalogs to the online public access catalog (OPAC) (Walker, 2009). While these tools have tended to increase the efficiency of search, some research indicates that there is a negative impact on the ability or inclination to browse since browsing behavior often involves perusal either of an entire work or resources located in nearby physical space (Prabha, Rice & Bunge, 1987 as summarized by Rorvig, 1998). A study by Sathe, Grady and Guise (2003) found differences in the preferences of print versus electronic versions of journals as well as differences in access and use. For example, print journals are used more for browse compared to electronic journals, including scanning of the table of contents. While electronic journals were favored for ease of searching and access, users of electronic journals found additional resources more through direct access to databases, while print journal readers were more likely to find additional references via browse behavior.

Although the OPAC has been the subject of research directed at restoring some browse capabilities to use of online systems (Beheshti, 1992), studies show that many students prefer to conduct both search and browse in popular databases such as Amazon and Google (Griffiths and Brophy, 2005; Yu and Young, 2004). Why this may be so can at least partially be explained by the observation that these popular sources create an environment that facilitates both search and browse behaviors, providing not only subject, term and full-indexed search but also providing highly visible recommendations and access to other resources based on social data and business analytics. These added

features creates an atmosphere conducive to what Bates (1989) calls “berrypicking,” the tendency of searchers to navigate research through a combination of alternating search (narrowing) and browse (broadening) behaviors.

A term often associated with browse behaviors in the literature is *serendipity*, meaning “an unsought, unintended, and/or unexpected discovery and/or learning experience that happens by accident and sagacity” (retrieved 5/1/2010 from <http://en.wiktionary.org/wiki/serendipity>). Serendipity combines the element of chance with the ability to recognize meaningful relations and associations. The relationship of serendipity to browse involves both access to facets of information about resources and sufficient motivation to fully capitalize on the curious or coincidental association of resources to each other as exposed by the facets.

Foster and Ford (2003, p. 323) quote Rice (1988) as saying, “...reference librarians know that what comes under the rubric of serendipity is often an actual, possibly subliminal, search strategy. Stated very generally, the potential for serendipity should be directly related to the number of different access points or potential ways of retrieving from a given system.”

While it may often be the case that a reader searches and finds a book out of specific cognitive need, it is also true that much book-seeking behavior, especially for pleasure or recreational reading, is browse oriented. To the extent that discovery has moved online, the limitations of browse in an online information-seeking environment is of considerable concern not only to the academic and research library community (Bui, 2012) but of information intermediaries across disciplines and communities. Social

gatekeeping is a framework for understanding how social networks can support and multiply the number of access points and potential ways of retrieving information from the online ecosystem. Serendipitous ties, those accidental yet critical links between individuals and between networks of social networks, extend the social gatekeeping framework beyond the traditional theories of diffusion of information and present a strategy for authors and publishers to increase online browse access points and information retrieval pathways.

### **Empirical Support for Social Gatekeeping and the Serendipitous Tie**

Social network analysis has contributed to an understanding of the means by which information diffuses through a social network. Granovetter (1973) used interview techniques to classify a social network according to tie strength and showed that information comes to a social network primarily through acquaintances of weak tie strength. Travers and Milgram (1969) engaged individuals to communicate a message through their network and tracked its path to a target individual. They showed that any two individuals are connected to each other by a chain of acquaintances and that information can travel from a source individual to any given target individual through an average of less than six information intermediaries. Gladwell (2000) suggested that Travers and Milgram's result could be explained by the existence of key individuals he called connectors whose influence spans multiple social networks and thus speeds up the propagation of information.

These experimental results and others like them are consistent with a gatekeeping framework, that is, individuals become aware of information, process it and then choose to pass it on to others or discard it. Gladwell's connectors are, in effect, social gatekeepers who control and intermediate the flow of information between and among acquaintances.

Most social network analysis data collection, as with the examples above, has traditionally been based on interviews or surveys that directly ask people about their connections and aspects of their relationships (Butts, 2008). These data collection techniques present challenges of recall, bias and error, and may not be effective at

uncovering discontinuous diffusion of information between individuals without measurable tie strength unless specifically designed to detect it. However computer mediated social networks generate large data sets that could be used for this purpose.

Abdessellem, Parris and Henderson (2012) discuss data collection techniques from online social networks and note the potential as well as some of the issues. One problem is that complete data sets are seldom available for research. They discuss the issue of privacy settings, for example, and observe that even though many individuals do not implement full privacy control, access to full and complete data sets on large social networks such as Facebook is not commonly available to most researchers. A few recent studies on very large data sets, however, provide some evidence for social gatekeeping and the serendipitous tie.

Bakshy et al (2012) was able to access Facebook's entire network of users internally to study information diffusion within the social network by analyzing the news feeds of approximately 253 million subjects in situ. The experiment was designed specifically to test assumptions about strong and weak ties in information propagation by studying the link and sharing patterns among connected individuals. Interestingly, exposure to (p. 521) "content that subjects may have seen through interfaces on Facebook other than what was posted on their feed" was perceived to be a threat to data quality, although this is the kind of information transfer that social gatekeeping proposes through serendipitous ties.

The experiment convincingly established that Granovetter's weak tie hypothesis and general assumptions about the impact of homophily on diffusion of information

noted in other research are supported – the probability of sharing was greatest when the information appeared on the feeds of homophilous individuals linked by strong ties, but the sheer number of weak tie relationships were responsible for the majority of information sharing on the site. The researchers conclude that simple contagion was an effective model for information diffusion and that (p. 526) “in large online environments, the low cost of disseminating information fosters diffusion dynamics that are different from situations where adoption is subject to positive externalities or carries a high cost.” This in itself offers support for a social gatekeeping framework, where the relatively low barrier to participation in computer mediated networks enables individual gatekeeping potential and the rapid and effective dissemination of information with a social network.

But in confirming diffusion of information within the network between individuals with either strong or weak tie relationships, the experimental design controlled for external influences including both the situation where friends act independently or situations where an individual is influenced to share through a communication in an external or independent communication channel, thus establishing a (p. 526) “lower bound on how much on-site sharing is due to interpersonal influence along any communication medium.” It might be expected that once information enters a social network, the great majority of information sharing occurs within the network among individuals with identifiable ties. The fact, however, that some degree of sharing occurs that is identifiable as not originating within the site or among individuals with explicit strong or weak ties suggests that the serendipitous tie may play a role in diffusion of information through a social network and in introducing information to a network.

Modifications of the experiment that might shed more light on the degree to which this occurs, and that might explicitly differentiate between internal and external information sources, could offer stronger confirmation of the degree of impact of the serendipitous tie. In effect, what the research controlled for as a confound could itself be a factor of interest.

Onnela et al (2007) analyzed a multi-million record dataset of mobile phone communications to determine the effect of strong and weak ties on network integrity. Unlike the Bakshy study, which was confined to a particular single social network, the dataset included all mobile phone calls of 20% of the total population of a country for a period of 18 weeks, thus spanning multiple individual social networks. The researchers assessed tie strength on the basis of the number and duration of calls between distinct numbers and examined the integrity of the network structure by selectively removing nodes represented by strong or weak ties until the network collapsed, that is, reached the point at which the structure of the network disintegrated. They observed, contrary to the expectation that strong ties form the basis of social networks, that removal of strong tie nodes resulted in a slight diminution of network structure, but did not cause the network to collapse. In contrast, removal of weak tie nodes caused a sudden and total collapse of the network structure. This study confirms Granovetter's weak tie hypothesis, however they also noted even weak tie connections have properties that tend to slow information flow. They conclude (p. 7336) "...we find that the observed coupling between the network structure and tie strengths significantly slows information flow, trapping it in communities, explaining why successful searches in social networks are conducted

primarily through intermediate- to weak-strength ties, while avoiding the hubs. Therefore to enhance the spreading of information, one needs to intentionally force it through the weak links, **or alternatively, adopt an active information search procedure** [emphasis added].” This suggests that information can become isolated within individual social communities and may not spread without resorting to a flow of information beyond that which occurs through weak ties.

Zhao, Wu and Xu (2010) examined the weak tie hypothesis using datasets from Facebook and YouTube in order to test the strategy of selecting strong ties, weak ties or randomly selected ties as a means to diffuse information sharing within the network. While the main findings confirmed the weak tie hypothesis, they found that specifically targeting weak ties for information sharing compared to random selection did not either speed up or more effectively distribute information through the network. These results suggest that while weak ties play an important role in information diffusion within a network, other factors come into play and that weak ties only partially mitigate issues such as information trapped within sub-communities.

None of these three studies explicitly attempted to explain information diffusion other than through strong or weak ties, but although all three serve to confirm generally the weak tie hypothesis, all three also strongly imply a certain amount of flow of information into and through social networks that is not accounted for by strong or weak tie relationships. For information that does not enter a network from an external source such as mass media, the diffusion of information not accounted for by strong or weak tie relationships can be accounted for by the exchange and transfer of information between

individuals without an identifiable tie relationship. These are the serendipitous ties, and their role in network information diffusion should be explored. Further research designed to look for evidence of the serendipitous tie would help quantify its influence and role in the diffusion of information.

### **Social Gatekeeping, Sales and the Review<sup>7</sup>**

Book reviews in mass media dates at least from the 17<sup>th</sup> century, emerging as a unique form of editorial in the early newspaper and periodical business following adoption of the printing press (Bry and Afferbach, 1961; McCutcheon, 1922). Book reviews eventually became important tools for newspapers to sell advertising; the first book reviews were published in the New York Times in 1896 as news articles (Lucey, 2006) and served as a focal point for advertisers interested in aligning products including books and other merchandise with interest in books and reading.

Reviews sell books (and most other products), and even negative reviews serve to inform readers about books and their particulars, resulting in increased sales (Sorensen and Rasmussen, 2004). Reviews help readers determine what is being published, how new books fit into existing literature, how they contribute to ongoing disciplinary dialog, and whether individual books deserve further attention (Ingram and Mills, 1989). Although negative ratings have more of a negative impact than positive ratings have a positive impact, both tend to increase sales compared to books with few or no reviews (Chevalier and Mayzlin, 2006; Hu, Liu & Zhang, 2008). Libraries in particular rely on reviews to inform purchases (Blake, 1989; Levine-Clark and Jobe, 2007; Serebnick, 1981; Serebnick and Cullars, 1980) and depend on a handful of trade review sources that review far fewer titles as a percentage of total titles available than even a decade ago.

---

<sup>7</sup> Portions of this section reflect Fulton 2010b, 2010c, 2009d, 2009e, unpublished

Currently, newspaper book reviews are in serious decline, with only a handful of newspapers now providing significant book review sections (Kurtz, 2009). Book reviews are among the features commonly eliminated as newspaper circulation figures decline (Grabois, 2007; Wasserman, 2007). Although numbers aren't readily available, it's reasonable to assume that popular and general circulation magazine book reviews follow suit as their circulation also continues to dwindle; between 2007 and 2009, at least 1,466 magazines ceased publication (Flamm, 2009). Traditional sources for the trades publish reviews of only a fraction of the available titles; *Publishers Weekly*, *Library Journal* and review arms of the American Library Association each review approximately 7,000 titles each per year with quite a bit of overlap (Wyatt, 2004).

Consumers therefore increasingly turn to online sources for discovery and to learn about the details of particular item. Much of this activity is socially driven. Major sources of consumer-driven information resources include social networks, blogs, user-contributed ratings and reviews on vendor sites and even videos (a recent query on ["book review"] on the video sharing site YouTube returned 379,000 results). According to a 2008 Pew Internet and American Life Survey, 66% of Internet users have bought something online, and 81% turn to the internet for research about products or services they are thinking of buying, with 20% of them doing so on a daily basis (Horrigan, 2008a). A 2007 Power Reviews survey found that 68% of online shoppers read at least four online reviews before making a purchase decision. Some surveys show that consumers place more confidence in consumer reviews than even expert reviews (Schmitt, 2007), in spite of the fact that online reviews are subject to manipulation and

outright falsification and fraud (Dellarocas, 2006; Harmon, 2004; Hu, Liu, Sambamurthy and Chen, 2008).

The contention here is that social gatekeeping functions similarly to the professional book review in informing readers of the existence of a title and of its particulars. Outputs of social gatekeeping in this context include not only consumer-produced book reviews specifically, but also commentary in blogs, social networks, comments on Web sites and various forms of face to face communication. Since professional reviews, and mass media in general, now cover only a fraction of the current book title output, social gatekeeping plays an increasingly critical role in bringing the long tail of book titles to the attention of the reader; the aspiring author or publisher must find a way to trigger social gatekeeping in order to make their work discoverable. The numbers don't have to be large to trigger information diffusion. In one survey, only 4% to 8% report having made an online review or comment, or posted a quality rating on internet purchases (Horrigan, 2008b), for example. Other surveys show that face to face word of mouth still accounts for a majority of information sharing and that most individuals are consumers rather than producers of information (Lang, 2012). These results tend to support the notion of the importance of social gatekeepers who act as persons of influence in the dissemination of information through social channels. Whether through mass media or social gatekeeping and the serendipitous tie, information has to penetrate a social network before word of mouth transmission and more passive methods of information diffusion can begin. Social gatekeepers who control information gates and the information made available for further dissemination start the process, and

information consumers help propagate information through more passive responses such as replying to an original post, sharing a message or participating in a computer-mediated network's "like" functionality.

## Popular Conceptions of Gatekeeping

Gatekeeping is a common metaphor in mainstream media, referring to the practices of publishers, agents, channels to market, review and literary criticism venues and the emerging power of both the author and the reader. Self-publishing generally, and more recently, e-books (both traditional and self-published), are often positioned in the popular media as disruptive to traditional gatekeepers and gatekeeping concepts. In particular, there has been much discussion about whether traditional gatekeeping is dead, and what might replace it. Much of the popular dialog reflects the conception of the gate as a chokepoint, that is, of information filtered out and gatekeepers standing between authors and readers.

A few of the memes from articles selected from recent newspapers, blogs and other popular media sources include these themes, headlines and quotes:

### *Publisher as Gatekeeper*

- EBook Leader Smashwords CEO: "We've Eliminated the Publisher as Gatekeeper" (Weir, 2011)
- The eBook Wars: "The argument against [traditional publishers] goes along many threads, all fueled by objections to publisher release delays of an eBook, the eBook's quality, and the price of the eBook, among others." (AmericanEditor, 2010)

### *Author as Gatekeeper*

- “For many years the choke point in publishing was distribution. That is no longer true with the rise of the eBook. So the traditional route of writer-agent-editor-publisher-sales forces-book buyer-bookstore-reader has been broken. We’ve got writer-reader (of course there is editing, formatting, etc. but that can be outsourced so it’s not a chokepoint any more). The real gatekeepers in publishing now? Authors. ... Most people would say readers are now the gatekeepers. To an extent they are. But here’s the deal: writers create the product. The quality of the product is going to determine how readers react to it. The ability to promote/market the product is going to determine if readers even get a chance to react to it.” (Mayer, 2011).

#### *Reader as Gatekeeper*

- “Book publishers see their role as gatekeepers shrink: Going through the gate still has certain benefits, but it's no longer the only way for authors to get to where they want to go. ... Godin, 50, said he realized that he no longer needed a publisher to distribute his work or to find an audience: He had cultivated a following of millions through his blog and speaking tours. ‘If an author has the choice of two distribution models, one that costs nothing and has no gatekeeper and the other has lots of gatekeepers and costs a lot of money, a lot of people will go with the free one,’ he said. ...Fewer and fewer people are walking into a bookstore. You have to

reach readers in other ways. Because, ultimately, the new gatekeepers will be the readers.” (Pham, 2010)

*Editors and Reviewers as Gatekeeper*

- “Editors and reviewers will remain, but their role will change, from gatekeepers to guides.” (Whitworth, 2009)

*Attribute as Gatekeeper*

- “Quality is the new gatekeeper. With a to-read list that’s cleared the 1000-book mark and is growing, I simply don’t have time to spend on anything less than a compelling, error-free and professional-quality book.” (Biba, 2011)

*Price as Gatekeeper*<sup>8</sup>

- “Books that have gone through the traditional gatekeeping role tend to support higher pricing than those that have not. I am willing to spend 99 cents for a nongatekept eBook because it is not much of an outlay — it’s like buying a lottery ticket; I am willing to gamble \$1 on odds of 6 million to 1 but I am not willing to pay \$5.99 for such an eBook because the risk of getting dreck is much too high. On the other hand, I am willing to spend

---

<sup>8</sup> Although the writer references publishers as gatekeepers in this quote, scripted filters can create automated gates to information thus providing a personalized approach to social gatekeeping based on selected criteria, in this case, price as filter.

\$7.99 for a gatekept eBook because the risk is generally that I will not enjoy the writer's style or I won't be in the mood for the particular genre, not that I will be stuck with dreck (although that, too, does happen and is happening with increasing frequency as the gatekeepers fumble around eBooks)." (AmericanEditor, 2011)

#### *Market as Gatekeeper*

- "This apparent anomaly of greater choice resulting in a narrower selection finds a corollary in Amazon's use of metrics to recommend titles based on previous purchases. The algorithms at work here are highly sophisticated and are widely credited with expanding consumer choice. Yet such metric-based systems can simultaneously increase the variety of books purchased by individual customers while decreasing the overall variety of books bought by everyone. This is because, as blogger Whimsley explains, "In Internet World the customers see further, but they are all looking out from the same tall hilltop." (Robinson, 2010)

#### *Librarian as Gatekeeper*

- "Today's librarian might indeed be an information management expert, but that's the revised job title, not the one HR knows about. The old corporate librarian was considered an archivist; the newer version is often called a gatekeeper or custodian, but I prefer to call them an information broker, an indexer and distributor of information. Gatekeeper sounds too

one way for me, though I get the quality maintenance role.” (Ericson, 2009)

### *Metadata as Gatekeeper*

- “Whoever controls the metadata controls the marketplace. This is because in a virtual economy, the metadata is just about the only way to convey marketing messages and prompt discovery. If OPACs evolve into virtual bookstores, a war for control of the metadata in OPACs will arise, as the metadata can then be linked to ecommerce capability. We should look for companies like Amazon and Google, not to mention OCLC, to attempt to insinuate themselves into OPACs.” (Esposito, 2011)

Two of the prevailing popular views of gatekeeping, that of editorial control and cultural preservation are explored in two works, Andrew Keen’s *Cult of the Amateur* (2007), and Steven Rosenbaum’s *Curation Nation* (2011).

Keen’s polemic against user generated content positions the read-write web as a threat to professionalism and high quality content traditionally filtered through old school mainstream gatekeepers such as publishers and editors, whose job it was to keep out the junk. He laments the demise of the traditional gatekeeper:

“Say good-bye to today’s experts and cultural gatekeepers – our reporters, news anchors, editors, music companies, and Hollywood movie studios. In today’s cult of the amateur, the monkeys are running the show. With their infinite typewriters, they are authoring the future. And we may not like how it reads.” (Kindle ed. loc. 117)

“Moreover, the free, user-generated content spawned and extolled by the Web 2.0 revolution is decimating the ranks of our cultural gatekeepers, as professional critics, journalists, editors, musicians, moviemakers and other purveyors of expert information are being replaced ... by amateur bloggers, hack reviewers, homespun moviemakers, and attic recording artists.” (Kindle ed. loc. 180)

But somewhat contradictorily, Keen acknowledges the continuing power of traditional gatekeepers even while positing their demise:

“The irony of a “democratized” media is that some content producers have more power than others. In a media without gatekeepers, where one’s real identity is often hidden or disguised, the truly empowered are the big companies with the huge advertising budgets. In theory, Web 2.0 gives amateurs a voice. But in reality, it’s often those with the loudest, most convincing message, and the most money to spread it, who are being heard.” (Kindle ed. loc 1060)

Rosenbaum, on the other hand, argues the other side of gatekeeping, the active selection of resources by people who filter the best, most relevant materials based on expert knowledge and subject familiarity. In Rosenbaum’s view, expert is not necessarily the credentialed professional, but the curation perspective of gatekeeping is that human filters pull the best, most relevant information from the information glut and package it for the information seeker:

“Curation is about something different than disintermediation. In fact, it’s about remediation. It’s about adding quality back into the equation and putting a human filter

between you and the overwhelming world of content abundance that is swirling around us every day.” (Kindle ed. loc. 262)

This is the information science perspective of gatekeeping, which Keen (via Rosenbaum) dismisses as democratic fiction:

“Curator is a euphemism for the NPR crowd who aren’t willing to utter ‘gatekeeper.’ That’s what a curator is. A curator doesn’t say yes most of the time; a curator says no. ...Maybe in NPR-speak, a curator says yes and a gatekeeper says no, but they’re doing the same thing.” (Kindle ed. loc. 1851).

Publishing expert John B. Thompson reflects on the concept of gatekeeping primarily from the narrow perspective of blocking or filtering content from passing through the gate (editorial gatekeeping, information selected out, saying “no”) while arguing at the same time that publishers are not “just” gatekeepers by virtue of their proactive role in content production (information filtered in, saying “yes”). He writes:

“In the first place, the notion of the gatekeeper is not really adequate as a way of characterizing the selective activities of editors and publishers. The idea of the gatekeeper suggests that there are authors queuing up to get through the gate, and the gatekeeper’s job is to decide who can go through and who will be turned away. This model may have been a reasonably accurate reflection of what happened in some sectors of the publishing industry decades ago, but it doesn’t bear much resemblance to the role of an editor in most publishing firms today.” (2005, p. 4)

“The publisher acts not just as a filter or gatekeeper but in many cases plays an active role in creating or conceiving a project, or in seeing the potential of something and helping the author bring it to fruition.” (2010, p. 19).

## **Multiple Regression**

Multiple Regression is a statistical tool most often used to form predictions about the behavior of a dependent variable based on changes in two or more independent variables, and as a tool to help understand the role of two or more independent variables on the variance observed in a dependent variable. It is widely used in a variety of disciplines in the social sciences including economics, education, sociology, communication, history and many others. Multiple regression is essentially based on correlation and as such cannot be used to prove causality. It can never be used as a tool of certainty in the proof of a significant theory. It has utility in providing supporting, if circumstantial, evidence for models of social systems, and can, if properly used and interpreted, provide some indication of the relative effect of multiple variables on dependent variables. However, it also can provide misleading results as a result of poor modeling or choice of variables (McDonald, 2013).

Rubinfeld (2011) notes that tests of hypothesis are appropriate for, among other types of investigations (pg. 319) "...cross-section analysis, where the data underlying the regression study have been chosen as a sample of a population at a particular point in time, and in a time-series analysis, when the data being evaluated cover a number of time periods." However, he notes several confounds that can make regression results unreliable or misleading. These include cases where the dependent variable may influence the independent variable (such as price and demand); correlation between independent variables, which can result in ascribing strength to one variable actually attributable to another; whether individual errors in the calculated model are actually

independent; the impact of outliers, which can greatly alter calculated results, and measurement error.

Manzi (2012) has been especially critical of non-experimental methods generally, and of regression in particular. Manzi strongly recommends the randomized field trial (RFT) as the gold standard for confirming the predictability and relative influence of independent variables on the behavior or characteristics of a dependent variable. With regards to regression, he notes that real-world models are invariably complex and involve a great number of factors that must either be accounted for or controlled. However, regression is sensitive to this issue in two ways. First, as the number of factors increases, the ability of regression to manage large numbers of independent variables with sufficient power decreases and the need for larger sample sizes increases. Second, if independent variables that may influence the dependent variables are omitted either by error or design or regression technique used, this omitted variable bias results in misleading weights given to the influence of the variables included in the model and can alter the significance of the model as a whole.

However Manzi does not claim that non-experimental methods have no place in the social science. Rather, he acknowledges that (Kindle Ed. loc. 2430) "...significant roles remain for other analytical methods [other than RFT]" and notes at least three use cases for non-experimental methods:

- "As an alternative to RFTs when they are not practical" (loc. 2437),

- “As a funnel that winnows out unpromising ideas at minimum expense before deploying the expensive tool of an RFT as the definitive arbiter of effectiveness...” (loc. 2474), and,
- “To develop hypotheses that can subsequently be rigorously tested.” (loc. 2482)

### **Search Engine Count as Dependent Variable**

The search engines Bing and Google (as well as some other not considered for this research such as Yahoo) are engineered to provide relevant results to information queries by returning a ranked list of URLs according to some propriety algorithm. Google Page Rank is a well-known system that operates on a massive database of the full-text indexing of all Web pages it elects to examine and estimates each site's importance and relevance according to all the other Websites that link to it. The technique was first explained by Brin and Page (1998) and has undergone considerable refinement since. Bing and other search engines use similar proprietary techniques. Along with the list of ranked links, the search engines provide an estimate of the count of the pages indexed in its database (the "hit count") that are relevant to the search (Estimated vs. Actual Number of Results, 2010).

It stands to reason that this data would be collected since the index of words, terms and phrases is part of the page rank algorithm. However, both Bing and Google provide estimates only to three significant figures (e.g. 123; 1,230; 12,300 etc.). So the figures cannot be treated as exact counts and it's not clear or transparent how the estimates provided to the public for a given search relate to actual page rank calculation. Because the data is derived from proprietary algorithms that cannot be independently validated and because the numbers are estimates instead of precise counts, the interpretation of the hit count is controversial among some researchers.

Google hit counts have been used in linguistic and phoneme analysis studies where word order and frequency are of interest, and there are some cautions for use of the

hit count depending on what is expected from the result. Kilgarriff (2007) addresses use of the hit count in the context of language analysis and find issues with the way Google handles parts of speech. He also notes the limit of 1,000 returns per query as problematic in terms of validation and suggests that an open source search engine could be developed that would better serve the academic community. However to date, none has been developed. Pollard (2007) notes these and other issues, such as the way Google normalizes words, does some transformation of diacritics and ligatures, and substitutes alternative spelling.

On the other hand, while noting some of the limitations and occasional inconsistencies of the hit count, Nakov and Hearst (2005) compared variability in n-gram counts across different search engines including Google, Yahoo and MSN (Bing's first incarnation) and found that variability over time was not sufficient to affect the interpretation of results of n-gram frequencies derived from the search engine results. Uyar (2009) found that the three search engines produced consistent but differing numbers of counts, with Google generally providing an upper limit estimate, Yahoo a lower limit and Live Search (Bing's immediate predecessor) somewhere in the middle, indicating that each of the search engines likely indexes a unique set of pages using differing estimation methods. However he concluded that all search engines produced reasonably precise results on single word queries with decreasing accuracy on multiple word queries. He did not address the issue of a quoted query on an exact phrase.

Janetzko (2008) has also investigated the validity of search engine hit counts, noting that in spite of some skepticism, search engine counts have gained acceptance as a

valid measurement tool for scientific research. He confirmed some concerns with inconsistent hit counts using Boolean constructions involving disjunction or negation, but concluded on the basis of extensive testing over a large number of parameters that hit count objectivity, reliability and validity were generally good.

Bagrow and ben\_Avraham (2005) used Google hit counts to examine the fame distribution of selected classes of scientists and other groups. Despite issues including inconsistencies with Boolean disjunction, the search engine hit count remains (p. 1) “an excellent tool for research.” The methodology generally confirmed prior studies and was based on the assumption that “scientists habitually use the [Web] as a professional means of communication and cite each other on the Web in relation to their published work.” However they also note that noise in the Google returns and available data precluded “sharp conclusions.”

For this research, Boolean constructs were not used. Two kinds of automated search scripts were run on Bing and Google. First, the Amazon ASIN (stock identification number for eBooks) entered as a single search term was searched weekly for 15 weeks on both Bing and Google. The ASIN is a unique identifying identifier consisting of 10 or more characters. The second search was on a quoted construction of author and title (e.g. [“The Litigators by John Grisham”]) which results in returns of high recall and precision, although the reliability of the hit count can't be explicitly verified.

The second quoted search was conducted weekly for 15 weeks on Bing and once in week 15 for Google.<sup>9</sup>

---

<sup>9</sup> At the time data was collected, Google aggressively blocked scripted access to search results based on quoted queries making weekly data collection tasks infeasible. Google data on this query was therefore collected once by hand towards the end of the data collection period. Details are documented in the Methodology chapter.

### **Amazon Sales Rank as Dependent Variable**

Amazon does not publish raw sales data for its products but does assign a sales rank to items such as eBooks. The sales rank represents the ranked order of sales according to a propriety formula, with books selling well appearing higher in the ranked list than those selling poorly. Thus, a title with a sales rank of 3,000 sells more books per unit of time at the time the sales rank is calculated than a book ranked 30,000, but the ranks don't provide the actual measure of sales. Use of this data in scientific research is subject to the same criticisms as the use of search engine hit count, namely that the number is proprietary, unverifiable and unreliable.

Several researchers, however, have independently deconstructed Amazon sales rank such that it can provide a useful tool for estimating sales from sales rank and for comparing estimated sales among groups of products (Brynjolfsson, Smith and Yu, 2003; Chevalier and Goolsbee 2003; Chevalier and Mayzlin, 2006; Dhanasobhon et al, 2007). In each case, investigators combined known sales data from independent sources with Amazon sales rank and then made purchases of selected titles to observe the impact on sales rank, which for popular books may be computed by Amazon as frequently as hourly and posted daily. The results of these investigations were consistent and demonstrated that book sales follow a standard Pareto distribution by which coefficients can be derived that yield close approximations of sales based on sales rank. The particulars of the coefficients vary over time as sales and Amazon's formula vary, so historical measurements such as these studies provided cannot be used to calculate sales at some later point in time. Some researchers have been content merely to use the inverse log of

sales rank for purposes of statistical analysis where a proxy for sales is needed (Hu, Liu, Bose & Shen, 2010). A few bloggers have attempted to track sales as a function of sales rank using current data and have arrived at basically the same conclusion. Rosenthal (2012) calculates a relatively recent analysis but suggests that the actual sales rank to sales numbers may be subject to change over time and that sales rank number should be collected over time to establish consistency. However while it may not be possible to determine actual sales with precision, the trends have been consistent over time so that techniques such as binning according to a logarithmic scale could be an effective statistical proxy for comparing sales.

## **CHAPTER 3 - METHODOLOGY**

### **Introduction to the Methodology**

The purposes of this research are to describe the use of social media by eBook authors, to investigate the extent to which author Web presence predicts increased eBook Web presence and sales, and to look non-exhaustively at selected books and authors for future areas of inquiry. The first portion of the research is descriptive and is designed to address questions concerning the nature and certain characteristics of a random sample of the current output of eBook titles and their authors, and to compare and contrast these characteristics with a sample of popular eBooks. The second portion of the research is designed to determine whether author Web presence and social media outreach is predictive of eBook Web presence and sales. The authors of a random and popular sample of current eBooks are categorized by their use of certain social media, and then the Web presence and sales of the titles in both random and popular samples are measured in regular intervals over time. Statistical tools are used to infer the relationship between author Web presence and book Web presence and sales. The final portion of the research is a non-exhaustive look at selected books and authors from the random and popular samples in order to further understand how eBook authors use social media and to suggest areas of inquiry for further study.

## **Explanation of Data Collection Techniques**

Several data gathering techniques were used to generate a random sample of eBook titles (the Random Sample) plus a list of popular titles (the Popular Sample) and to collect and organize data about them and about their authors for use with the research. These techniques include direct scripted access to a provider's database through a programming interface, HTML parsing of web pages, and manual search and data collection techniques.

Developer and affiliate programs are offered by Amazon, Bing, Goodreads and other Web platforms. The purpose of these programs is to provide access to certain information primarily for the purpose of creating third-party applications that add value to the platform. All operate similarly by issuing a private security key that allows direct programmatic access to selected data they choose to make available through an Applications Program Interface (API), typically using standard data interchange formats such as XML or JSON (JavaScript Object Notation). These data can be retrieved using a scripting language such as PHP, Perl or Python and then extracted to databases and spreadsheets for analysis.

The data available through a provider's API is usually limited. Where desired information is available on a given web page but not available through the API, it is possible to script automated collection of that data through a technique known as HTML parsing, sometimes also called screen-scraping. Screen scraping is part art and part

programming. The technique requires manual examination of the HTML of representative Web pages including, if applicable, the underlying referrer pages.<sup>10</sup>

Once the appropriate pages are retrieved, the HTML contents are manually examined to determine identification of the location and offsets of the desired data. For example, the raw HTML of a retrieved web page may contain the following string of text somewhere within hundreds of lines of HTML code:

```
<a class="title" href="http://www.amazon.com/Acts-of-Citizenship-  
ebook/dp/B00BZ75AUW/ref=sr_1_2?s=books&ie=UTF8&qid=13651  
13657&sr=1-2">Acts of Citizenship</a>
```

This indicates a hyperlink (bolded text) pointing to a book with the title *Acts of Citizenship*. The hyperlink embeds the Amazon stock number at the end of it, terminated by the text beginning with /ref. A script can be written to locate the text `<a class="title" href="` from within all the text on the page, extract just the characters following up to but not including /ref and store the extracted text string into a database for further action.

Once these locations and markers are identified and the scripting written and tested, the HTML can be retrieved using a computer script and an input file of URLs to be examined, such as the set returned by a browse command or a prepared data file. The data can then be parsed for extraction to databases and spreadsheets for analysis.

---

<sup>10</sup> Referrer pages are used internally by the site to extract a batch of results and display them in the client's browser using AJAX (Asynchronous JavaScript and XML) techniques. A referrer page is the URL a site uses to pull the next page of information for a query result and is not necessarily the URL that displays when a user clicks the next page link.

Some data was collected manually, and even manual data collection can be partially scripted. In addition to scripting with HTML parsing, it is also possible to create a spreadsheet macro command that automates the loading of a web page based on a URL stored in a cell. This makes visual examination of web pages relatively quick.

Finally, much of the data related to author Web presence was obtained through manual search techniques including general purpose search engines and the search functionality built into various social media sites.

## **Technology Environment**

All fully scripted data collection was managed through a VMWare Workstation-hosted Linux Virtual Machine installed on a Windows 7 PC. The installed software included the LAMP stack consisting of Apache server, MySQL and PHP, plus the PHPMyAdmin graphical management console. This facilitated the running of PHP scripts using both APIs and HTML parsing. The raw data was collected into MySQL using PHP and then exported to Excel spreadsheets and Microsoft Access database tables using the PhPMyAdmin console. The scripts described in sections that follow were developed and tested between fall of 2011 and spring of 2012.

Manual data collection was collected via Microsoft Excel and browsers using the researcher's laptop and home network connectivity or computers and networks located in the University of Arizona main library or branches of the Tucson Public Library.

## **Description and Collection of the Random and Popular Samples**

Data collection commenced with generation of a random sample of eBook titles and retrieval of a non-random sample of Amazon's most popular eBook titles, which are described and categorized as indicated in the sections that follow.

The first set, randomly generated from a population of all eBooks published and released over the first week of April of 2012, is hereafter referred to as the *Random Sample*.

The second set, retrieved on the same day as the Random Sample from a list of most popular titles as reported by Amazon, is hereafter referred to as the *Popular Sample*. The Popular Sample reflects Amazon's most popular titles on a given day as reported by Amazon and was not randomly collected or organized.

### *Random Sample*

The general process of using a representative random sample to classify and categorize books was introduced by Bradley, Fulton, and Helm (2012). They identified a universe of self-published books numbering around 385,000 as of 2008 and devised techniques to derive a random sample of 348 for analysis.

The parameters of the current universe of eBooks are unknown. In April of 2013, Amazon stocked approximately two million Kindle titles as reported by the result count of their advanced search function. Barnes and Nobel's search function indicates over three million available titles on a blank search for Nook Books (their proprietary eBook format). Many of these are undoubtedly duplicated, but both sites hold unique titles, so

there is no way to determine even among these two what the number of eBook titles really is through these retailers, let alone through a multiplicity of other sources. Further, there does not appear to be a method that could be used to retrieve a complete list.

Browse queries on Amazon limit the number of returns, and in many cases, the number of returns is less than the total number of titles matching a general query. This generally precludes using browse as a mechanism to retrieve an entire population of titles. However, further investigation by this researcher on Amazon resulted in the discovery that while most browse requests limit returns to 100 pages of results or fewer, with 12 results per page, one particular browse URL would return up to 400 pages each of all fiction and non-fiction eBook titles published in the last 30 days from date of query. Further, the results could be reverse-chronologically ordered by release date thus eliminating Amazon's non-random proprietary ordering such as "ordered by relevance." Amazon's release rate as of March, 2012 was approximately 23,000 each in the fiction and non-fiction categories, with browse returns of 12 items per page of results, so while the complete monthly output could not be retrieved, a complete list of several date-ordered consecutive days of new title output could be collected.

The underlying URL schema for queries that could be used to automate a scripted retrieval of approximately 4,800 fiction and 4,800 non-fiction books published in the last 30 days in reverse chronological order was retrieved and manipulated to determine the critical components:

```
http://www.amazon.com/gp/search/ref=sr_nr_n_0?rh=n%3A133140011%2Cn%3A%21133141011%2Cn%3A154606011%2Cp_n_date%3A1249100011%2Cn%
```

3A157028011&**page=1**&bbn=154606011&**sort=daterank**&ie=UTF8&qid=1330  
631568

The important parameters are indicated in bold, where the **ref** parameter ends in 0 for fiction or 1 for non-fiction, **page** equals the specific browse page of the query and **sort** requests the results in date rank order.<sup>11</sup> A two-step PHP script to cycle through pages 1 – 400 of each query and use HTML parsing was executed to extract the Amazon ASIN stock number of each eBook returned by the browse.

The script was run on April 6, 2012, and it retrieved a list of all titles with Amazon release dates ranging from March 30, 2012 through April 6, 2012. After elimination of duplicates<sup>12</sup> and elimination of titles with a release date of the partial days March 30 and April 6, a total of 8,525 records representing the complete population of books released on Amazon between March 31, 2012 and April 5, 2012 inclusive was imported into an Access database and numbered sequentially. Next, the random number generator on the website random.org, which uses atmospheric noise to generate true random numbers, was used to produce 500 random numbers between 1 and 8,525. The list included 16 duplicate numbers leaving 484 numbers that were matched to numbered titles and selected as the Random Sample, providing a confidence interval (margin of error) for the entire Random Sample of +/- 4.33 percent at a 95% confidence level. One

---

<sup>11</sup> Amazon changes URL parameters from time to time, so this query may no longer be current.

<sup>12</sup> Some of the referrer pages were found to contain previews of the next page in sequence.

title was pulled or unpublished from Amazon prior to data collection leaving a Random Sample total population of N=483.

### *Popular Sample*

A second dataset consisting of 200 of Amazon's most popular eBooks was retrieved on April 6, 2012 (the same date as the first collection of the Random Sample data) from a list of Amazon's most popular titles, which Amazon maintains and updates as often as hourly. The list of titles was retrieved through HTML parsing techniques directly from the Amazon Web site. Of these, 100 were top paid downloads, for which sales rank was listed, and 100 were the top free downloaded eBooks on that day, for which no specific sales rank was listed.

The overwhelming majority of free titles in the Popular Sample were actually titles for which a price was usually charged. It is an increasingly common practice among publishers, and especially self-publishers, to offer a title for a very limited time at no cost in order to boost rank, downloads and visibility, on the theory that these factors will result in larger volumes of sales once the free download period expires. In fact, all but 13 of the initially free titles went back to paid status during the data collection period, with the only volumes remaining free being public domain classics. Once the books' free status expired, sales rank and offer price were reported and collected.

The Popular Sample was drawn only once, on April 6, 2012, and then tracked weekly along with the Random Sample. Over the course of data collection, a number of titles in the Popular Sample were put on free download status for a limited period of time.

The Popular Sample therefore includes books that were free on April 6 but which reverted to paid status, and books that were on paid status on April 6, but which may have been offered for free at some point during the data collection period. The fluid environment of paid and free titles offers an opportunity for further investigation, but for this research, no further differentiation was attempted in the analyses that follow in Chapter 4, which was based on their initial status at time of collection as free or paid.

Some of the items in the Popular Sample as retrieved were actually electronic games, and a few were listed in a way the scraping routine didn't pick up. The total count of the Popular Sample after removal of these items was  $N=190$ .

It is important to reiterate that the Popular Sample is not a random sample nor is it intended or used as a control group. It was collected and tracked along with, but separately from, the Random Sample dataset in order to provide exemplars that could be used to compare aspects of popular eBooks and authors with eBooks and authors in the Random Sample. Unlike the Random Sample, which consists of eBooks released during one short period of time, publication dates of the Popular Sample were not date limited and so may have been in the market for months or years, and undergone previous cycles of popularity. Further, author Web presence may have changed over the course of the release where the title was on the market for an extended period of time. These factors should be taken into account, therefore, when considering effect size and relative strength of the predictor variables for the regressions run on the Popular Sample compared to the Random Sample.

## **Preparation of the Datasets**

Once the Amazon stock numbers (ASINs) for the Random and Popular Samples were scraped from the browse queries, the Amazon Affiliate API was used in conjunction with PHP automated scripting to pull selected static information about each eBook and extract it to databases and spreadsheets for processing. These are metadata that would not be expected to change over time and include full author and title, publisher, list price, complete URL of the eBook, subject classification terms, and ISBN, if one was assigned.

This information was used, in turn, to generate a list of URLs, authors and titles necessary for collection of the data associated with the dependent and independent variables. The dependent variables are those associated with book Web presence and sales and were collected weekly over 15 weeks for a total of sixteen measurements. The independent variables are those associated with author Web presence. These data were collected in one batch through manual iterative searching during a period at the beginning of data collection and took approximately four weeks to complete. These processes and the specific data that was collected are described in further detail in the sections that follow.

## **Dependent Variable Weekly Data Collection**

### *Strategy and Approach*

The dependent variables measured for this research include eBook Web presence as measured by search engine hit count, and sales as measured by Amazon sales rank. These data were collected weekly each Friday morning beginning April 6, 2012, and concluding on Friday, July 20, 2012, for a total of 16 measurements spanning 15 weeks.

### Amazon Data

Amazon sales rank figures at a given point in time were retrieved using the Amazon Affiliate API to pull the information from Amazon's database. Information not retrievable from the API was scraped from each of the respective eBook Web pages and included the number of reviews in each of the review bins (1- through 5-star) and the offer price, which can vary from list price. Prepared data files consisting of the Amazon stock numbers of the Random and Popular datasets were read by an automated PHP script that retrieved both the API data and the data from the eBook Web page using HTML parsing.

### Search Engine Data

Two search strategies were developed to measure eBook Web presence. No single query or multiple queries could be constructed that would return orthogonally every web page pertaining to a specific eBook with high precision and recall. Search engines are designed to return the most relevant results but not an exhaustive set of relevant results.

Nevertheless, queries about eBooks can be constructed that, while not exhaustive, are high in both precision and recall.

The Amazon ASIN is a 10 digit alphanumeric identifier that functions similarly to a book ISBN. eBooks are not always assigned an ISBN, however all eBooks sold by Amazon are assigned a unique ASIN. Queries on the ASIN therefore produce results with high precision and recall<sup>13</sup> and can serve as a relative indicator of Web presence, with higher numbers of returns reflecting greater diffusion of eBook metadata on the Web.

The second strategy involved searching on a specific quoted string containing the title and author. While queries with multiple search terms might produce returns with high relevance listed early in the result set, the entire result set often returns some results that reflect only one or some of the terms. For example, a query such as [John Grisham The Litigators] will return the desired book early in the result set, but the result set may include pages relevant only to litigators, or to Grishams other than John, or to other matches according to the search engine's proprietary algorithms, and those would be reflected in the total search engine hits.

Quoted phrase queries, however, return results of only those URLs that match the quoted phrase exactly.<sup>14</sup> For example, a query such as ["The Litigators by John

---

<sup>13</sup> In theory, another organization could adopt a similar naming convention and produce results that duplicate the Amazon ASIN, but none were detected during the course of examining returns for this research. The odds that any given 10-character alphanumeric string might match an Amazon ASIN purely by chance would be 1 in 3.6<sup>11</sup>.

<sup>14</sup> Google holds US patent US 7222299 relating to the detection of quoted text within a document. Bing uses its own undisclosed proprietary algorithms.

Grisham”] will return only pages containing that phrase<sup>15</sup>. This particular literal form of the phrase [“*title by author*”] is very commonly used in reviews, commentary and book descriptions and serves a purpose similar to the ASIN to represent a query with high precision and recall whose count results reflect relative Web diffusion. Weekly data collection of these dependent variables, therefore, included, in part, automated search engine queries on a quoted string of title and author (e.g. query: [“The Litigators by John Grisham”]).

Two search engines were selected for data collection through scripted query, Bing and Google. Text files were prepared with title, author and Amazon ASIN information that could be processed by a PHP script.

For Bing, a developer key was used to access search data returned in JSON format from the Bing search API. Data collected included search engine hit counts for the quoted title-author search and the ASIN for each eBook in the Random and Popular Samples. For Google, an HTML parsing technique was developed since Google had suspended developer access to the search API.

#### *Dependent Variables - Weekly Data Collection*

Once the Random Sample and Popular Sample were finalized, the dependent variables of book Web presence and Sales were retrieved weekly for 15 weeks according to the methods described above. As collected, the author and title data received from the

---

<sup>15</sup> On quoted phrase searches, Google does not do stemming on words and also includes stop words otherwise ignored such as *by*, *the*, *a*, and so on. (Blachman, 2012).

Amazon API, which also appear on the eBook's web page, were inconsistent and required manual normalization in order to create a consistent source data file for weekly data collection. Frequently, title fields were supplemented with series, language or other information that needed to be stripped. In other cases, certain punctuation and non-text queries were found to produce erratic results. Some selected examples of norming include:

Table 1 - Author/Title Normalization

<b>Amazon Title Field</b>	<b>Manually Normed</b>	<b>Reason</b>
Between "I Will" & "I Do"	Between I Will I Do	Search engines ignore or misinterpret certain punctuation and symbols
The Empty House and Other Ghost Stories (Fully Illustrated)(Annotated)	The Empty House and Other Ghost Stories	Remove descriptive text not part of the title
Destiny Kills (6th Sam Casey Mystery)	Destiny Kills	Remove series
Tamed by a Texan (Harlequin American Romance)	Tamed by a Texan	Remove series

<b>Amazon Title Field</b>	<b>Manually Normed</b>	<b>Reason</b>
Discovery, The	The Discovery	Normalize title

Norming the author field consisted primarily of assuring that all entries were in order of first name last name (e.g. John Grisham rather than Grisham, John). In some cases, author names appeared differently in the Amazon database than in the eBook itself, in which case, the eBook usage prevailed. Approximately one-third of the Random Sample and less than that of the Popular Sample required some norming prior to weekly data collection.

Some issues were discovered in succeeding weeks as eBooks were manually examined. This resulted in re-norming either the author or title of 13 eBooks in the Random Sample and 8 books in the Popular Sample to assure that the final counts were as accurate as possible. All re-norming took place prior to collection of the final three data counts used for the regression analyses (see Chapter 4). Future work involving time series data across the entire data collection period would need to take these changes into account, or eliminate the re-normed books from the Random and Popular Samples.

Weekly data collection using automated scripting then commenced on April 6, 2012 and ran for 15 weeks through July 20, 2012 providing 16 counts for each variable collected. Both Amazon and Bing<sup>16</sup> automated routines worked as tested.

Google proved more problematic. At the time this data collection was conducted, Google had discontinued access to its University Research Program for Google Search which provided API access to search results. At the same time, they began aggressive blocking of automated or semi-automated queries of certain kinds including scripted queries containing quoted phrases. Queries originating from scripts such as PHP scripts were completely blocked. Further, the IP addresses of computers sending semi-automated queries on quoted phrases using techniques such as spreadsheet macros were blocked after 75 or so returns for up to several hours, thus limiting the amount of data that could be retrieved using that method.

However, Google did not block scripted retrieval of searches based on the ASIN and also continued to provide scripted queries on the ASIN limited to returns from Blogs. Therefore, weekly data collection on ASIN from Google remained feasible. In addition, a single semi-automated search on the quoted title-author phrase was conducted manually in the last week of data collection using multiple computers at libraries. Although this precluded computing a three period moving weighted average for the Google quoted

---

<sup>16</sup> Bing stopped including the search engine hit count in its API effective 8/1/2012 shortly after the data collection phase of this research was completed.

phrase search (see Chapter 4), data was made available for a final count that could be used with the regression.

The following table lists the dependent variable data that were collected beginning in April, 2012, and the frequency:

Table 2 - Data Collection, Dependent Variables

<b>Data</b>	<b>Collected</b>
Amazon Sales Rank	Weekly
Amazon Offer Price	Weekly
Amazon # Reviews 1-star through 5-star	Weekly
Bing Author-Title	Weekly
Bing ASIN	Weekly
Google ASIN – General	Weekly
Google ASIN – Blogs	Weekly
Google Author-Title	Once

Note: Amazon offer price – what a title is selling for at any given point in time – is subject to change, and many titles are made available for a limited time at no charge as

a special promotion. This technique is thought to boost sales by increasing exposure due to social interest in the period following the promotion. When a title is offered at zero price during a promotion, Amazon does not report a sales rank. Collection of the offer price data allowed identification of those cases of missing sales rank data as opposed to cases indicating an actual sales rank of zero, which indicates no sales have been recorded for the item.

## Collection and Classification of Independent (Predictor) Variables

### *Strategy and Approach*

Determination of the independent (predictor)<sup>17</sup> variables relating to author Web presence involved iterative and recursive manual searching. The basic information that was collected and coded included determining whether the author maintained a Web page, a Blog, a Facebook page, a Twitter account or had author presence in GoodReads and Amazon.<sup>18</sup>

Sites were classified as Web pages if they contained only static or dynamic one-way (author to reader) content that did not include reader comment, discussion and feedback capability. Sites were characterized as Blog sites if they originated on a blog platform using traditional blog style sheets (e.g. Wordpress, Blogspot) or if all pages consisted of reverse chronologically ordered postings with reader comment and feedback. Sites that contained both one-way and two-way blog-style communication pages were classified as both Blog and Web site even if a common base URL was used.

---

<sup>17</sup> Independent variables use in regression are often referred to as predictor variables because they are used to predict observed variance in the dependent variable tested.

<sup>18</sup> Authors with books in the Amazon marketplace, either traditionally published or self-published, may register and claim an author page that provides access to internal Amazon sales data and also the ability to post a biography, trade reviews and links to other social media sites. Goodreads provides similar functionality, but in addition often supplements an unregistered author page with biographical and social media information even if the author doesn't personally claim the page and provide the information. As a result, Goodreads-produced author pages frequently contain as much or more biographical and social information as a page actually claimed by an author.

For Facebook pages, number of “likes” or “friends” were captured as well as whether the page was a personal page or business page. For Twitter accounts, the number of tweets, numbers of accounts followed and numbers of accounts following were collected.

#### *Collection of the Independent variables*

Independent variable data on author Web presence (Amazon author page, Goodreads author page, Web page, Blog, Facebook page and Twitter account) plus other descriptive data began with scripted retrieval of the basic bibliographic and sales information from the Amazon API database on April 6, 2012. This information was used to prepare Excel worksheets for manual search and retrieval of the independent variables and to take a closer look at various aspects of the Web pages identified.

The first pass included reviewing each eBook page on the Amazon site to see if the author had established an author page. Authors registering with Amazon may optionally include a description/biography and links to a Facebook, Blog or Twitter account from their page; where available, that information was collected and verified.

Next, search engines were used to query on author name. For many authors, a link to a website or blog was retrievable within the first three pages of the search engine result set. If the first search was unsuccessful or if it produced large numbers of irrelevant or ambiguous results (for example if the name was a common one), the author’s name was combined with the term *author*, then combined with the title of the book, and then combined with the publisher name. When a Web page or blog was discovered and

confirmed, a review of the contents also frequently included a link to other social media pages such as Facebook, Twitter or the author's Goodreads page; if so, the information was collected and verified.

Twitter provides an internal search engine function to locate accounts by name or key words. If the author name was insufficient to locate a match, the title of the book was used. In most cases, the displayed biographical summary was sufficient to verify a match. For common names, the results could number in the hundreds or thousands without an obvious match. Twitter results were returned first ordered by relevance and then by activity. For large returns, the results were reviewed until the activity for displayed accounts showed zero or one for tweet activity. For matches that did not include sufficient information to verify the account (e.g. that the John Smith identified was actually the John Smith that wrote the particular book), tweets were examined for content that reflected the title or reflected titles of other books by the desired author. For verified matches, the biographical summary also frequently contained information leading to or helping to verify Web pages or Facebook accounts. For Twitter accounts additional data retrieved included number of tweets, number of accounts following the author, and the number of accounts the author follows.

A similar strategy was used to search Facebook. Facebook returns consist either of personal pages or business pages. For personal pages, the number of friends was captured; for business pages, the number of likes was captured.

The social reading site Goodreads.com was searched both on title and author. Authors who register with the site are listed as a Goodreads Author on relevant pages and

may then create a biography and communicate on the site with fans and other authors.

Author information may also be added to author pages by Goodreads staff and volunteers. As with the other social sites, author pages frequently included links to other social media.

Data collection for the most of the independent (predictor) variables took approximately four weeks of iterative searching and verification, with some continued clean up and verification proceeding through the data collection period ending July 20, 2012.

The following table summarizes independent (predictor) variable data that was collected beginning April 6, 2012:

Table 3 - Data Collection, Independent Variables

<b>Variable</b>	<b>Variable Type</b>
Amazon Author	Yes / No (coded 1/0)
Goodreads Author Page	Yes / No (coded 1/0)
Facebook Page	Yes / No (coded 1/0); Friend/like (count)
Twitter Page	Yes / No (coded 1/0); Tweets (count); Followers (count); Following (count)
Web Site	Yes / No (coded 1/0)
Blog	Yes / No (coded 1/0)

### **Classification of Authors, Titles and Publishers**

In addition to the specifics of dependent and independent variables described above, a third step in preparing data for analysis consisted of classifying certain characteristics of the authors and eBooks in the Random and Popular Samples. Examples include subject/genre classification, price, publication model (self- or mainstream published), number of other editions (if any), and other factors. Certain characteristics of authors were also classified, such as number of other books and eBooks published. Classification serves multiple purposes in the context of this research. The first portion of the Chapter 4 summarizes selected key characteristics of the current state of eBook title production drawing from both the Random and Popular Samples. This establishes a frame of reference for the quantitative analysis that follows and creates a baseline of data that informs future work. Selected aspects of the classification also inform and provide a basis for a more fine-grained look at the quantitative analysis and supports like to like comparisons. For example, author Web presence may impact book Web presence and sales differently for works of fiction vs. works of non-fiction, as suggested by the research questions in Chapter 1.

Self-published books seldom are labeled explicitly as such and determining whether a book is self-published or published by a traditional mainstream publisher or one of the estimated 80,000 to 90,000 small and niche US publishers presents challenges. For purposes of this research, some rules and procedures were used to classify a book as self-published.

- If a title's publisher of record in the Amazon database or on the title's Amazon page was blank or listed as Amazon Digital Services, the book was presumed to be self-published. These are typically eBooks uploaded directly to the Kindle site for publication without publisher information, which commercial publishers would be unlikely to do. These were often found to be paired with a print on demand version labeled with CreateSpace as the publisher, Amazon's imprint for self-published authors.
- If the Amazon database listed the author's name in the publisher field, the book was presumed to be self-published.
- If the Amazon record (or print version of the title, if applicable and available) listed a known self-publishing author services firm such as CreateSpace, Lulu, Authorhouse, the title was presumed to be self-published.
- For other books, the publisher was searched and reviewed to determine if the author self-published either completely independently or through a fee-based author services company with a chosen an imprint (publisher) name other than the author's own name as a publisher. If the publisher web site listed only the author, or the author and one other party as published authors, the title was presumed to be self-published. This accounts for self-published individuals who work with co-authors or illustrators, but not publishing firms representing three or more authors. This rule is in basic harmony with the Library of Congress Cataloging in Print program that requires, among other things, at

least three published authors represented by a publishing company to qualify as non-self-published and eligible for LOC pre-publication cataloging service.

## **Data Analysis and Statistical Approach**

### *Strategy and Approach*

The second phase of data analysis included primarily the use of multiple regression to determine whether and to what extent the independent variables predict variance in the dependent variables; in other words, whether and to what extent author Web presence predicts book Web presence and sales. The data collected for the Random and Popular Sample meet the basic requirements of the regression method, namely: the dependent variables are interval (sales in number of books; search engine hit counts, number of reviews), the predictor variables are nominal and dichotomous (yes/no coded 1/0) or interval (Facebook number of likes, Twitter number of tweets), and there are sufficient cases for the number of predictor variables to be tested (for most tests, up to six predictor variables with between 100 and 400 expected cases).

SPSS was used for all statistical analysis. Excel and Access were used to organize data for analysis.

### *Statistical Setup and Preparation*

Three categories of dependent variable data were collected for the Random and Popular Samples, book Web presence, book sales and consumer review count.

Book Web presence was measured by search engine hit counts on the Amazon stock number and also by a quoted author title search. The Google search engine was used to query the ASIN number using standard search and also separately with return results limited to blog sites. The Bing search engine was used to query both the ASIN

number and the author title quoted search. These data were collected weekly for 15 weeks using automated scripted routines. In addition, the Google search engine was queried manually at the end of week 15 on the author title quoted search.

Book sales were calculated by extrapolating estimated sales from Amazon sales rank data, which was collected weekly. The relationship between Amazon sales rank and sales has been experimentally determined to model a power law of the general form:

Equation 1 – Sales as a Function of Sales Rank

$$\text{Sales} = a * \text{SalesRank}^n$$

(Brynjolfsson et al, 2003; Chevalier and Goolsbee, 2003; Fenner, Levene & Loizou, 2010, Iba et al, 2008; Pinto et al, 2012)

Actual sales rank by sales data estimated from graphs provided by Rosenthal (2012) for weekly book sales was entered into SPSS for a power curve estimation.

The results estimated the parameters of Equation 1 to be:  $a = 1,044,389.9$  and  $n = -1.078$ , with  $R \text{ Squared} = .878$ ,  $F_{1,5} = 36.142$ ,  $p = .002$ . The exponent value  $n$ , which defines the shape of the power curve, is consistent with estimates from the research noted above.

With the exception of the manual Google search on author title for which only one data point was collected for each case, the dependent variables used for the regressions that follow were calculated as the moving weighted average of the final 3 weekly counts using a standard calculation with 3 period smoothing, namely, given smoothing period  $n=3$  and measurements  $x_i$ ,  $x_{i-1}$  and  $x_{i-2}$ :

## Equation 2 – Moving Weighted Average Formula

$$X_{ma3} = (x_i * n + x_{i-1} * (n-1) + x_{i-2} * (n-2)) / (n * (n+1)) / 2$$

For example, given counts at the end of weeks 13, 14 and 15 of 100 (count 14), 125 (count 15) and 150 (count 16):

## Equation 3 - Moving Weighted Average Example

$$DV = (150 * 3 + 125 * 2 + 100) / 6 = 133$$

Use of the weighted moving average smooths spikes and noise in the data and is often used in analysis of time series data such as analysis of stock prices or merchandise sold over time, where data is subject to a certain amount of noise and where it is desirable to place the most weight on the latest value.<sup>19</sup> In the case of search engine counts, the results provided by the search engines are estimates which may include noise from week to week as a result of the total number of indexes used by the search engine for a given estimate at a particular point in time. For sales, the Amazon sales rank for slow moving titles is calculated less frequently than for fast moving titles inducing some noise in the reported sales rank of slow sellers merely as a result of shifts in sales of the more popular books. The weighted moving average gives the most weight to the latest count – desired in this case under the presumption that at the early stage in the book cycle, both sales and

---

<sup>19</sup> Exponential weighted average is also commonly used to smooth noise in data but tends to give somewhat more weight to earlier results. As a practical matter, there is little difference between WMA and EMA on three data points as used here. In the example shown, the EMA calculates to 131.25 vs. 133 for WMA. The WMA at the 16<sup>th</sup> week would be the same even if the entire 16 terms in the series were calculated since the WMA only looks back as far as the smoothing period. For most cases, the EMA would result in a lower average if all 16 terms were calculated.

diffusion are trending up – and helps reduce the impact of noise as a hidden source of variance in the regressions.

As collected, the dependent variable raw data were highly skewed. This is expected since as noted above, book sales vs. sales rank are experimentally known to fit a power law distribution and have been found to be log linear (Brynjolfsson et al, 2003). Visual examination of the search engine hit count data closely resembled the sales data and a curve estimation of Google search engine hit count vs. rank order confirmed close fit to a power law curve  $y = ax^n$  where  $a = 61,322.027$ ,  $n = -1.604$  with  $R^2 = .903$ ,  $F(1,323) = 3005.582$ ,  $p < .001$ . Therefore, all dependent variables were transformed prior to using them in the regression analysis in order to meet the assumption of normality as nearly as possible.

A number of different transforms are used to normalize non-normal data for analysis (e.g. log, square root, reciprocal and other transforms), and as a practical matter, one may choose the method that provides the best results, since the data are not fundamentally changed with respect to each other but simply transformed by a common factor or process (Howell, 2007). The fact that book sales fit a power series curve suggests that a log transform might provide the best results.

The Box-Cox transformation is a form of logarithmic transform that uses iterative estimation to transform the data such that the standardized mean is at or near zero, the standard deviation is as close as possible to one, and skew and kurtosis are as close to zero as possible (Box & Cox, 1964). Although full normality cannot be guaranteed, Box-Cox often results in a transformation that approximates normality to a reasonable degree

(NIST/SEMATECH, 2012; Sakia, 1992). Recent versions of SPSS include an algorithm to compute the Box-Cox transformation, which was used on all the dependent variable data prior to computing the regression in order to reasonably approximate normality in both the dependent variables and the plotted residuals.

Six categories of independent (predictor) variables reflecting author Web presence were collected as dichotomous data coded 0 for no presence and 1 for presence of an Amazon author page, a Goodreads author page, a Web site, a blog, a Facebook page and a Twitter account. Each dependent variable was then tested against the predictor variables using the Enter method<sup>20</sup> of regression in SPSS.

---

<sup>20</sup> SPSS offers different ways of entering predictor variables for analysis. The simplest method is the Enter method, where all variables are entered at once. For some types of regression models, results can be further refined by entering variables in stages, such as stepwise, forwards and backwards. For this analysis, only the Enter method was used.

*Independent Variable Frequency Tables*

Table 4 - Distribution of the predictor variables from the Random Sample

Predictor Variable Statistics							
N=325		Amazon Author Page	Good-reads Page	Blog	Web	Face-book	Twitter acct.
N	Valid	325	325	325	325	325	325
	Freq = 0	226	187	272	184	215	202
	Freq = 1	99	138	53	141	110	123
	% = 0	69.5	57.5	83.7	56.6	66.2	62.2
	% = 1	30.5	42.5	16.3	43.4	33.8	37.8

Table 5 - Distribution of the predictor variables for the Popular Sample

Statistics							
N=179		Amazon Author Page	Good-reads Page	Blog	Web	Face-book	Twitter acct.
N	Valid	179	179	179	179	179	179
	Freq = 0	25	5	103	26	26	57
	Freq = 1	154	174	76	153	153	122
	% = 0	14	2.8	57.5	14.5	14.5	31.8
	% = 1	86	97.2	42.5	85.5	85.5	68.2

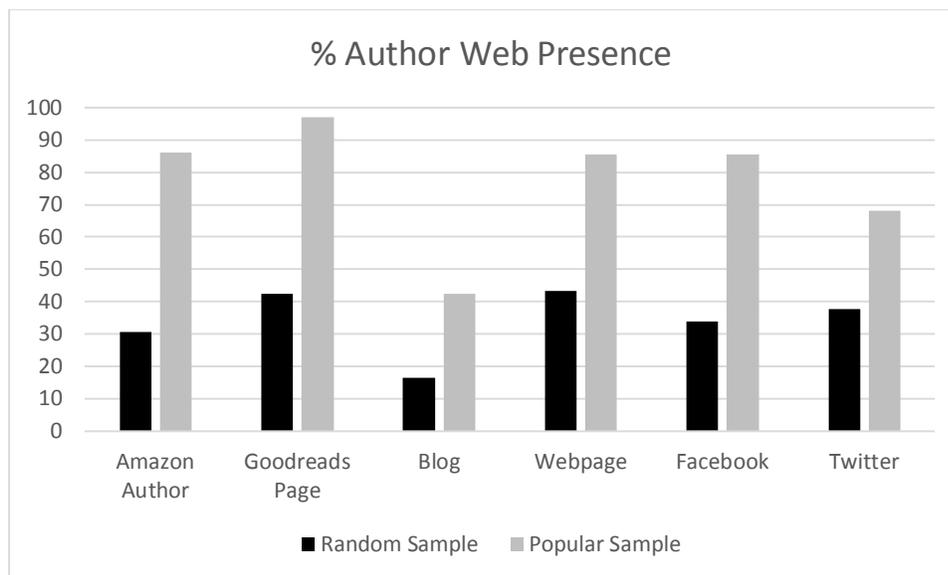


Figure 3- Author Web Presence

The figure above shows that author Web presence is considerably greater for the Popular Sample than for the Random Sample.

#### *Contamination of the Dependent Variable*

There is a potential for contamination of the dependent variable (DV) counts on ASIN and Author-Title (but not sales or reviews) because a web search on a particular search value (ASIN or specific quote author-title search phrase) might include that value on the author's web site or blog if it contained the search value. Thus, the count for DV could include some number of Web pages also counted as independent variables (IVs) to the extent that the ASIN or exact quoted author title search phrase appeared on the author's Web site or blog. However, not all author Web or blog data pages necessarily contain either the Amazon ASIN or the exact quoted author-title search phrase.

To determine whether this was a problem, a manual search was conducted in the final week of data collection that restricted the search to the author's Web or blog site. The true value of the DV, should contamination be a significant problem, would be the sum of the site-specific hit counts subtracted from the general search hit counts, e.g. for the ASIN search

Equation 4 - DV Contamination Formula 1

$$DV_{\text{adjusted\_count}} = \text{count}[\text{ASIN}] - \text{count}[\text{ASIN site:authorweb}] - \text{count}[\text{ASIN site:authorblog}]$$

However, if the log-transformed difference used in the regression computations is small, that is, if:

Equation 5 - DV Contamination Equation 2<sup>21</sup>

$$\text{Log}(\text{DVcount}) \approx \text{Log}(\text{DVadjusted})$$

Then, the contamination will have little practical significance on the calculated result of the regression. The data show that was the case.

For the Random Sample ASIN search on Google, 356 of 361 (99% of N=361) cases were uncontaminated, that is, the ASIN was not found on a search of the author's Web site or blog, if present, or the author didn't have a blog or website. Of the 5 records where the search on the author's blog and web site were positive, the contamination was

---

<sup>21</sup> The log is shown because the regression computations use log-transformed DVs.

between 1% and 9% of total count. The average change to the log-transformed dependent variable would be - 0.022, or - 0.43% on only the five cases.

For the Random Sample Author-Title search on Google, 351 (97% of N=361) cases were uncontaminated. Of the 10 records where the search on the author's blog and Web site were positive, the contamination, with the exception of a single outlier at 20% (1 out of 5 total hits) was between <0.01% and 0.3% of the total count. The average correction to the log-transformed dependent variable would be -.008 or - 0.229% on only the ten cases.

A higher number of cases in the Popular Sample were contaminated, primarily because considerably more Popular Sample authors maintain blog and or Web sites and the titles in the Popular Sample were on the market for a longer period of time. However the considerably greater number of search engine hit counts on popular authors and titles serves to dilute the impact of the contamination to near zero levels.

For the Popular Sample ASIN search on Google, 172 (95% of N=182) cases were uncontaminated, that is, the ASIN was not found on a search of the author's Web site or blog, if present, or the author didn't have a blog or Web site. Of the 10 cases where the search on the author's blog and web site were positive, the contamination was between 0.03% and 1.1% of total count. The average correction to the log-transformed dependent variable would be - 0.002, or - 0.053% on only the 10 cases.

For the Popular Sample author-title search on Google, 122 (67% of N=182) cases were uncontaminated, that is, the ASIN was not found on a search of the author's Web

site or blog, if present, or the author didn't have a blog or website. Of the 60 records where the search on the author's blog and web site were positive, the contamination was between  $<.001\%$  and  $1.4\%$  of total count. This shows that the count of contamination is very much smaller than the total count. The average correction to the log-transformed dependent variable would be  $-0.0002$ , or  $-0.004\%$ .

Similar results were observed in the Bing searches. Only one contaminated case was found in the Random Sample in each of the ASIN and author-title searches. In the Popular Sample, twelve contaminated records were found in the ASIN search. In the author-title search, 72 cases of contamination were observed but with near zero average corrections to the log-transformed dependent variable due to the high counts of the search results, as was found with the Google examples.

These results show that the issue of potential contamination of the DVs is of no practical consequence in the calculation or interpretation of the regressions.

**Ethical Considerations**

Because some of the research involves review of messages, commentary and other writing posted by individuals on the Web, in social networks and elsewhere, the protocol for the research was submitted to the University of Arizona Human Subjects Independent Review Board for a determination of its classification as human research. Because only publicly posted information will be subject to data collection and no private or personally protected information will be reviewed or collected, the research protocol was ruled not to meet the definition of human research, effective April 5, 2012, and exempt from any further action, reporting or record keeping with respect to the Independent Review Board.

## CHAPTER 4 - RESULTS

### **Phase I – Description, Frequencies and Central Tendencies of the Random and Popular Samples.**

#### *Research Questions and Hypotheses – Phase I*

Selected characteristics of the data sets are summarized in the subsections that follow. Phase I of the research addresses Research Questions 1 and 2:

RQ1: What comprises the current output of eBook production? What are its characteristics and alternative formats?

RQ2: How do the characteristics of the current output of ebooks break out by subcategory including genre and subject, length and price, self-published vs. mainstream published, and other factors?

Upon examination, characteristics of certain titles rendered them unsuitable for the regression in Phase II of the research (e.g. public domain reprints, magazines released as eBooks and other aspects). These filters and their effects on the Random and Popular Samples are described in the sections that follow.

Because the Popular Sample is not a random sample, care should be exercised in the interpretation of the following data. The Random Sample represents analysis of the output of eBooks published and released on Amazon during the first week in April, 2012; the Popular Sample represents analysis of the most popular books bought and/or

downloaded from Amazon on April 6, 2012 according to a list developed and published by Amazon.

The Random Sample represents what authors and publishers are releasing to the reading public. Since it was randomly selected, it is representative of the population of eBooks from which it was drawn, subject to chance variation and the margin of error.

The Popular Sample consists of the most popular eBooks that readers are buying and/or downloading, as determined by Amazon. It is not a random selection, and due to the way Amazon constructs the list, it may be biased along multiple dimensions. Therefore, frequency and central tendency data are not representative of eBooks generally available on Amazon or elsewhere and reflect only the characteristics of the Popular Sample. The primary purpose of the Popular Sample for this research is to serve as a collection of exemplars of eBooks that have been discovered and are most popular at a given point in time among Amazon's customer base. It serves as a point of comparison to the characteristics of the Random Sample and the pool from which the Random Sample was drawn.

*Language*

A total of 58 titles (12% of N=483) in the Random Sample were written in languages other than English, as follows:

Table 6 - Foreign Language Counts

<b>Language</b>	<b>Count</b>
Albanian	1
Catalan	1
Dutch	1
French	17
German	12
Greek	1
Italian	5
Portuguese	2
Spanish	16
Swahili	2

These titles were removed from the pool leaving 425 titles in the Random Sample selected for further analysis.

All of the titles in the Popular Sample were written in English.

### *Reprint Jungle*

Manual review of each of the remaining titles revealed that 62 titles in the Random Sample (14.6% of N=425) were comprised of the category of books classified as “reprint jungle” by Bradley, Fulton and Helm (2012). These titles include primarily books in the public domain reprinted with new edition information and copyright dates, along with private label book spam (non-unique material licensed to be reprinted under the licensee’s name). While of interest to the general description of the Random Sample, these books were excluded from the primary research questions involving author Web presence, for obvious reasons.

In the Popular Sample, 8 (4% of N=190) fell into the reprint jungle category and were excluded from further examination of author title properties.

Observation: The reprint jungle remains an issue of even slightly greater magnitude than first analyzed by Bradley, Fulton and Helm (2012) on a 2008 sample of printed self-published works. These titles will continue to cause bibliographic and consumer confusion.

### *Other Adjustments, Initial*

During the first two to three weeks of data collection, three titles from the Random Sample were removed at least temporarily from the Amazon market place, making it impossible to collect basic data. These titles were removed from further analysis, leaving an initial Random Sample of titles considered for analysis on author and

book Web presence of N=360. Additional titles dropped off by the end of the data collection period, analyzed further in a section following.

No titles were initially unpublished or removed from the Popular Sample, although 3 were removed at a later point in time.

Observation: the ease with which books may be published, unpublished and republished is a cause of potential concern for bibliographic records

### *Self-Published Books*

In the Random Sample, following the classification rules described in the previous chapter, 262 titles (73% of N=360) were classified as self-published.

Of the Popular Sample, 71 titles (39% of N=182) were classified as self-published. Of these, 61 titles (67% of N=91) were from the portion of the list initially offered as free and 10 titles (11% of N=91) were from the portion of the list originally listed as Paid.

Observation: Only 13 titles in the Popular Sample remained free during the entire data collection period. Therefore, this statistic is probably best interpreted as indicating that self-published authors are more inclined to experiment with free trials than traditional publishers. Because the Amazon list of popular titles included many titles offered for free download at the time of collection, the Popular Sample (but not the Random Sample) is probably biased with respect to the number of self-published titles. Since the Popular Sample was not randomly selected, the numbers of self-published titles in the Popular Sample can't be used to generalize to the whole population of eBooks or

even the number of self-published books on Amazon. The Random Sample, on the other hand, is generalizable at least as far as eBooks published on Amazon around the time the Random Sample was generated.

Fee-based author services publishers provide self-publishing authors with a variety of services on an author-pays business model. Frequently, these publishing companies will also act as publisher of record and provide an ISBN registration from their pool of numbers. Of the self-published titles in the Random Sample, 29 (11% of N=262) listed fee-based author services publishers as the publisher of record. Of the Popular Sample, 5 titles (7% of N=71) were identified with fee-based author services companies listed as publisher of record. Note that this doesn't mean that the remainder of the self-published titles didn't use fee-based author services. Many fee-based author services publishers will allow the author to designate their own imprint as publisher.

Observation: self-publishing dominates the Amazon kindle market in terms of title production (but not necessarily of sales) at present. Although mainstream titles are increasingly available in eBook format and make up an increasingly large share of market for mainstream publishers, the sheer output of self-published titles makes discovery of self-published titles problematic for authors.

### *General Availability*

Barnes and Noble was searched by author and title to determine if other editions of the Amazon titles were available on that site, indicating the author was seeking wider availability.

143 (40% of N=360) of titles in the Random Sample were found also listed in the Barnes and Noble catalog, with 217 (60% of N=360) remaining Amazon exclusives, at the time data was collected during the first four weeks of the data collection period.

In contrast, 157 titles (86% of N=182) in the Popular Sample were found also listed in the Barnes and Noble catalog, with 25 (14% of N=182) remaining Amazon exclusives.

Observation: Best-selling titles are usually available in multiple formats in multiple market channels.

#### *Print Availability*

The Amazon record was searched to determine if a print version of the eBook title was available. In the Random Sample, 336 (80% of N=422) were eBook only, with 86 (20% of N=422) available on Amazon in print versions. Of the 336 eBook only titles, 234 (70% of N=336) were self-published.

In contrast, in the Popular Sample, 49 (25% of N=190) were eBook only, with 141 (75% of N=190) available on Amazon in print versions. Of the 49 titles in eBook only format, 30 were self-published.

Observation: Self-published authors are more likely to publish in eBook format exclusively than mainstream publishers, who make titles available in a variety of formats. This may be due to the expense and technical expertise necessary to produce books in multiple formats, or it may reflect preferential treatment by Amazon for authors who agree to exclusivity or some combination of factors.

### *Miscellaneous Attributes*

In the course of examining the Amazon page for each book, observations were made on miscellaneous attributes.

#### Description

Complete metadata has been shown to associate positively with sales, including information such as a complete description (Breedt & Walter, 2012). In the Random Sample, 43 (10% of N=423) lacked a description on the Amazon page. Of the remaining 90% that included a description, 14 (3.7% of N=380) included or referenced a review.

In contrast, only 3 (1.5% of N=189) in the Popular Sample lacked a description, and all three of those were reprint Classics. Of the 186 that included a description, 85 (46% of N=186) included or referenced a review.

Observation: Reviews, even negative ones, associate positively with sales (Sorensen and Rasmussen, 2004); self-published authors may still lack access to mass media editorial gatekeepers. Reader-contributed reviews and social media engagement may or may not compensate for the lack of trade reviews.

#### Kindle Loan Program

Kindle offers publishers the option of participating in an eBook loan program as a way of promoting their book. In exchange for a period of exclusivity on Amazon or other considerations, publishers receive royalties on loaned eBooks. Of the Random Sample,

110 (26% of N=423) participated in the loan program. Of the Popular Sample, 71 (37.5% of N=189) participated in the program

#### Enhanced / Optimized

Amazon lists some titles as either enhanced eBooks (containing links, multimedia, or other features) or optimized for larger screens and/or tablets, presumably due to images or graphics. In the Random sample, none were listed as enhanced but 13 (3% of N=423) were listed as optimized, usually indicating color graphics or fine detail such as charts and graphs. Only 1 book in the Popular Sample was listed as optimized and none as enhanced.

#### Subject Terms

Amazon does not use standard classification schemes for subject indexing and instead has developed their own system of terms and combinations of terms. For the Random Sample, a total of 392 separate subject terms were used to describe the titles. The 10 most used terms, in decreasing order are:

- Literature & Fiction
- Erotica
- Action & Adventure
- Contemporary
- Romance
- Education

- Contemporary Fiction
- Historical
- Short Stories
- Poetry

For the Popular Sample, a total of 172 subject terms were used to describe the titles. The 10 most used terms, in decreasing order are:

- Literature & Fiction
- Contemporary
- Suspense
- Romance
- Contemporary Fiction
- Historical
- Romantic Suspense
- Women Sleuths
- Mystery
- Christian Living

Observation: A complete and thorough subject analysis would require a different approach to collection of subject terms. Since Amazon does not use a standard classification system such as Dewey or BISAC, subject analysis is challenging, and further complicated by the question of who assigns the subject headings. In the case of self-published titles, these may be author supplied. Mainstream publishers can take

advantage of the Library of Congress' Cataloging in Print program and are more likely to employ professionals to supply subject metadata. This is an area for future research.

### eBook Erotica

While the Romance genre often includes some explicit sexual content, a number of eBooks in the Random and Popular Samples are either labeled erotica or reference explicit content in the book descriptions. Some media critics have pointed to eBook erotica as an emerging literary trend (Ledbetter, 2010; Bosman, 2012) In the Random Sample, 36 (8.5% of N=423) include the term *erotica* as a subject key word. Of the Popular Sample, none include the term; however the Popular Sample includes all three books in the *50 Shades of Gray* trilogy and *Jude Outlaw* by Jan Springer, who self describes as an erotic romance author on her Website. At the time of data collection, these were categorized with key subject terms *Contemporary Romance*. The *50 Shades* series has since been updated with an erotica subtitle, although the Springer novel has not at the time of this writing.

### *Final Adjustments to the Samples*

In addition to the adjustments previously noted (elimination of foreign language and reprint jungle titles), some titles were found unsuitable for the author Web presence regression analysis. These included titles that were actually releases of magazines, some federal documents and a small number of anthologies or collections with multiple authors. Once these were eliminated, the counts suitable for the regression analysis using author Web presence as independent variables stood at 349 for the Random Sample and

182 for the Popular Sample. Prior to the completion of data collection, 24 of the Random Sample were unpublished or removed from Amazon leaving a final count of 325. For the Popular Sample, 3 were unpublished or removed leaving a final count of 179.

#### *Sales – Random Sample*

All titles in the Random Sample were offered for sale at a price at least once during the data collection period; that is, none were permanent free downloads. Of 322 in the final Random Sample with complete data for all periods, 76 (24% of N=322) never acquired a sales rank indicating that there were no Amazon sales by the 16<sup>th</sup> week after publication. Of these, 68 (89% of N=76) were self-published. Viewed another way, 68 (29% of N=235) self-published books recorded no sales by the 16<sup>th</sup> week after publishing. In contrast, only 8 (9% of N=87) traditionally published books were unsold by the 16<sup>th</sup> week of publication.

Observation: These data point to the problem of sales for self-published books and to the need for discovery strategies.

#### *Sales as a Function of Social Media Outreach – Random Sample*

Another way to look at sales is grouped by use of social media. Of 328 cases in the Random Sample with data, 99 (30% of N=328) used none of the six categories of social media outreach measured (Amazon author page, Goodreads author page, Facebook account, Twitter account, Web site and blog). Of the 229 cases that did use social media outreach (70% of N=328), the average number of categories used was 2.2 with a median of 2.

The percentage of cases in the Random Sample with no recorded sales during the data collection period was 35% for the group that did not use social media vs. 19% for the group that did use at least one or more categories of social media.

From the Random Sample, the median sales rank for the group that did not use social media and that sold at least one book was 499,751. The median sales rank for the group that used at least one or more categories of social media and that sold at least one book was 328,465. When converted from sales rank to sales, this translates to a 36% increase in the median number of books sold per period by authors who use at least one category of social media outreach.

Observation: These data strongly indicate that author participation in social media does result in at least a modest sales increase. This also offers support for the significant results of the regression study as described below in the Phase II research.

#### *Price Manipulation – Random Sample*

Manipulation of selling price is a popularly promoted method of bringing attention to a title. Of 322 titles in the Random Sample with complete offer price information for all periods, 116 (36% of N=322) lowered prices at least once during the period of data collection. Most of these, 77 (66% of N=116) were self-published titles. In all, 36 (11% of N=322) dropped the price to \$0.00 for at least one period and of these, 32 (89% of N=36) were self-published.

Observation: Price manipulation strategies are used much more often by self-published authors than by mainstream publishers.

*Pricing – Random Sample*

For self-published books, the central tendencies of minimum and maximum offer prices are (units in \$):

Table 7 - Pricing, Self-published Books

		min	max
N	Valid	235	235
	Missing	0	0
Mean		3.57	4.43
Median		2.99	2.99
Mode		.99	2.99

For mainstream published books, the central tendencies are (units in \$):

Table 8 - Pricing, Traditionally Published Books

		min	max
N	Valid	87	87
	Missing	0	0
Mean		7.7492	8.9400
Median		6.4000	7.9900
Mode		9.99	9.99

In both subsets of the Random Sample (mainstream published and self-published) one or two high-priced outliers were responsible for the mean exceeding the median.

Observation: the mode of \$9.99 for traditionally published books reflects Amazon's pricing policy in effect at the time the Random Sample was collected. Since then, retail prices have demonstrated greater volatility as publishers and retailers have battled in federal court over pricing models.

## **Phase II – Quantitative Analyses**

### *Purpose*

The purpose of the second phase of data analysis was to provide evidence that would support or controvert the hypotheses associating author Web presence with book Web presence and sales. These hypotheses directly address the issue of discovery and sales. The results of the investigation provide empirical support for Social Gatekeeping, which suggests an explanatory mechanism and framework by which author and publisher engagement with social media increases visibility in social networks.

### *Primary Regression Models*

In the following sections, key statistics from the regression analyses are presented. For each of the seven dependent variables, a regression was computed using the six independent variables as predictors.

$R^2$ , the coefficient of determination, represents the effect size of the model, that is, the percent of variance in the dependent variable predicted by the independent variables. For example, an  $R^2$  of .25 would mean that the independent (predictor) variables, as a group, account for 25% of the observed variance in the dependent variable.

The F statistic and p value indicate whether the regression model is statistically significant, that is the odds that the variance observed might have occurred by chance. For this research, results were considered significant if there was less than a 5% chance ( $p < .05$ ) that the observed variance was due to chance. If the p value of the F statistic is

calculated to be  $\geq .05$ , then the null hypothesis, that the model is not predictive of the dependent variable, is not rejected.

The standardized Beta, sometimes also called the standardized coefficient, is a measure of the influence of each of the individual predictor variables on the dependent variable, assuming the remainder of the independent variables were held constant, expressed in terms of standard deviations. For example, a standardized Beta of .35 indicates that for each increase of one standard deviation of the predictor variable, the average increase (rate of change) in the dependent variable will be 35% of the conditional mean of the dependent variable, assuming the other independent variables in the model are held constant (Multiple Linear Regression Model, 2011). Betas can be either positive or negative, depending on the direction of influence. Each of the Betas is also tested for significance, again using the 5% chance ( $p < .05$ ) criterion.

Each regression model is sensitive to all variables included, and the  $R^2$  values reflect the effect size of the complete model including all predictor variables. The Beta of each of the independent variables does not explain the total effect on the dependent variable as if it were computed independently. Rather, each Beta represents the combined effect of adding that independent variable to the model, if the effects of all other variables in the model are already accounted for. Therefore, each coefficient may change when other variables are added to or deleted from the model (Interpreting Regression Coefficients, 2013).

Where the significance of the Beta for an individual independent variable is not significant, that is, where  $p \geq .05$ , the likelihood that the individual predictor variable

contributes to the model by chance is greater. If a predictor variable is removed from the regression, the  $R^2$  value might change, because as noted, the effect on  $R^2$  of each of the independent variables represents a combined effect. But if the Beta for the predictor was not significant, that means that predictor variable is more likely to have contributed to the model by chance and no statement can be made about the effect of that independent variable on the dependent variable. Another way of saying this is that “no statistically significant linear dependence of the mean of the dependent variable on the independent variable was detected” (Multiple Linear Regression Model, 2011). For this reason, Betas for non-significant independent variables are not reported in the sections that follow.

For purposes of this research, only the single regression using SPSS’ “Enter” method was performed; further attempts to refine the model in terms of individual predictors were not undertaken. Therefore, where each model is significant, it provides support for the general statement that author Web presence is predictive of book Web presence, sales or reviews. The p value of the independent variable Betas indicate the odds that the contribution of the independent variable occurred by chance. Where  $p < .05$ , the results are considered significant for purposes of this research. Where  $p \geq .05$ , the null hypothesis that the independent variable has no predictive value, cannot be rejected.

In the discussions that follow, only the statistics listed above are reported. In all, 14 regressions (7 for the Random Sample and 7 for the Popular Sample) were computed. A summary observation and table follow each of these two sections.

A more complete summary of each of the regressions is presented in Appendix A and includes results of the tests of normality and other statistics, and a review of the residuals.

*Regressions computed on the Random Sample*

DV = Google search engine counts on ASIN, Random Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .181$ ,  $F(6,318) = 11.711$ ,  $p < .001$ . Three of the independent variables had significant Betas as follows:

Table 9 - Google ASIN Betas, Random Sample

Predictor	Standardized Beta	Significance
Amazon Author Page	.185	.001
Goodreads Author Page	.182	.002
Facebook Page	.154	.015

Standardized Betas for Web page, Blog and Twitter account were not significant.

Put another way, the regression is significant for author Web presence associating positively with Google search engine counts on the Amazon Stock Identification Number. An Amazon author page, Goodreads page and Facebook page were significant predictors. The contribution of Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Google search engine counts, ASIN in Blog Pages, Random Sample

For the Random Sample, the frequency table of calculated weighted moving averages of the dependent variable showed that nearly 95% of cases in the Random Sample had 1 or fewer average counts, meaning the search found almost no penetration of the ASIN in the blog space as reported by Google. Although regression reported a significant model, with very low  $R^2$ , neither the transformed dependent variable nor the residuals were deemed normal or sufficient enough to consider the results reliable and so are not reported here.

DV = Google search engine counts on Author - Title, Random Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .132$ ,  $F(6,318) = 8.060$ ,  $p < .001$ .

Only the Goodreads predictor had a significant standardized beta of .288,  $p < .001$

Put another way, the regression is significant for author Web presence associating positively with Google search engine counts on the Author / Title search. Only the Goodreads page was a significant predictor. The contributions of Amazon page, Web page, Facebook, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Bing search engine counts on ASIN, Random Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Bing search engine hit counts, with  $R^2 = .167$ ,  $F(6,318) = 10.625$ ,  $p < .001$ . Three of the independent variables had significant Betas as follows:

Table 10 - Bing ASIN Betas, Random Sample

Predictor	Standardized Beta	Significance
Amazon Author Page	.114	.038
Goodreads Author Page	.290	.000
Facebook Page	.150	.019

Put another way, the regression is significant for author Web presence associating positively with Bing search engine counts on the Amazon Stock Identification Number. An Amazon author page, Goodreads page and Facebook page were significant predictors. The contribution of Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Bing search engine counts on Author – Title, Random Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Bing search engine hit counts, with  $R^2 = .379$ ,  $F(6,318) = 32.333$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Table 11 - Bing A/T Betas, Random Sample

Predictor	Standardized Beta	Significance
Goodreads Author Page	.444	.000
Facebook Page	.161	.003

Put another way, the regression is significant for author Web presence associating positively with Bing search engine counts on Author - Title. A Goodreads page and Facebook page were significant predictors. The contribution of Amazon author page, Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Amazon Sales, Random Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Amazon sales, with  $R^2 = .255$ ,  $F(6,315) = 17.959$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Table 12 - Amazon Sales Betas, Random Sample

Predictor	Standardized Beta	Significance
Amazon Author Page	.135	.010
Goodreads Author Page	.352	.000

Put another way, the regression is significant for author Web presence associating positively with Sales on Amazon. An Amazon author page and a Goodreads page were significant predictors. The contribution of Web page, Facebook, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Amazon Review Count, Random Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Amazon review count, with  $R^2 = .272$ ,  $F(6,315) = 19.639$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Table 13 - Amazon Reviews Betas, Random Sample

Predictor	Standardized Beta	Significance
Goodreads page	.385	.000
Facebook page	.225	.000

Put another way, the regression is significant for author Web presence associating positively with Amazon review count. A Goodreads page and Facebook page were significant predictors. The contribution of Amazon page, Web page, Facebook, blog and twitter account failed to achieve significance at  $p < .05$ .

*Summary – Random Sample Regressions*

The following table summarizes the significant predictors for each of the regressions computed on the Random Sample.

Table 14 - Summary of Significant Predictors, Random Sample

RANDOM SAMPLE		SIGNIFICANT PREDICTORS					
Source	Type	Amazon Author Page	Goodreads Page	Web	Blog	Facebook	Twitter
Google	ASIN	X	X			X	
Google	Author/ Title		X				
Bing	ASIN	X	X			X	
Bing	Author/ Title		X			X	
Amazon	Sales	X	X				
Amazon	Reviews		X			X	

Observations: There is consistency across the regressions computed for each of the dependent variables in the Random Sample. Both the Bing and Google searches on ASIN were predicted by Amazon author page, Goodreads page and Facebook page. The Goodreads page was predictive of both the Bing and Google searches on author and title. The Goodreads page was predictive for all dependent variables and demonstrated the highest Betas.

Of particular note is that all six regression models computed on the Random Sample were significant and consistent with the hypotheses that author Web presence and participation in social media associates positively with book Web presence, sales and reviews.

*Regressions Computed on the Popular Sample*

DV = Google search engine counts on ASIN, Popular Sample

Using the Enter method, after removal of an obvious outlier, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .170$ ,  $F(6,171) = 5.847$ ,  $p < .001$ . Three of the independent variables had significant Betas as follows:

Table 15 - Google ASIN Betas, Popular Sample

Predictor	Standardized Beta	Significance
Goodreads Author Page	.100	.047
Web Page	.162	.037
Facebook Page	.243	.002

Put another way, the regression is significant for author Web presence associating positively with Google search engine counts on the Amazon Stock Identification Number. A Goodreads page, Web page and Facebook page were significant predictors. The contribution of Amazon page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Google search engine counts, ASIN in Blog Pages, Popular Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .283$ ,  $F(6,161) = 2.479$ ,  $p = .025$ .

Of the predictor variables, after removal of an outlier, only Facebook had a significant Beta of .204,  $p=.015$

Put another way, the regression is significant for author Web presence associating positively with Google search engine counts on the Amazon Stock Identification Number in blogs. A Facebook page was the only significant predictor. The contribution of Amazon author page, Goodreads page, Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Google search engine counts on Author - Title, Popular Sample

Using the Enter method, after removal of an outlier, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .119$ ,  $F(6,161) = 3.839$ ,  $p = .001$ . Two of the independent variables had significant Betas as follows:

Table 16 - Google A/T Betas, Popular Sample

Predictor	Standardized Beta	Significance
Web page	.167	.036
Facebook	.185	.024

Put another way, the regression is significant for author Web presence associating positively with Google search engine counts on the author – title search. A Web page and Facebook page were significant predictors. The contribution of Amazon Page, Goodreads page, Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Bing search engine counts on ASIN, Popular Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Bing search engine hit counts, with  $R^2 = .112$ ,  $F(6,172) = 3.633$ ,  $p = .002$ . Two of the independent variables had significant Betas as follows:

Table 17 - Bing ASIN Betas, Popular Sample

Predictor	Standardized Beta	Significance
Web Page	.168	.035
Facebook	.206	.012

Put another way, the regression is significant for author Web presence associating positively with Bing search engine counts on the Amazon Stock Identification Number. A Web page and Facebook page were significant predictors. The contribution of Amazon author page, Goodreads page, Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Bing search engine counts on Author – Title, Popular Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Bing search engine hit counts, with  $R^2 = .191$ ,  $F(6,172) = 6.757$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Table 18 - Bing A/T, Popular Sample

Predictor	Standardized Beta	Significance
Web page	.161	.034
Facebook Page	.302	.000

Put another way, the regression is significant for author Web presence associating positively with Bing search engine counts on the author – title search. A Web page and Facebook page were significant predictors. The contribution of Amazon author page, Goodreads page, Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Amazon Sales, Popular Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Amazon sales, with  $R^2 = .156$ ,  $F(6,171) = 5.273$ ,  $p < .001$ . One of the independent variables had a significant Beta as follows:

Table 19 - Amazon Sales Betas, Popular Sample

Predictor	Standardized Beta	Significance
Facebook	.296	.000

Put another way, the regression is significant for author Web presence associating positively with Amazon sales. A Facebook page was the only significant predictor. The contribution of Amazon author page, Goodreads page, Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

DV = Amazon Review Count, Popular Sample

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Amazon review count, with  $R^2 = .197$ ,  $F(6,171) = 7.012$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Table 20 - Amazon Reviews Betas, Popular Sample

Predictor	Standardized Beta	Significance
Facebook page	.295	.000
Amazon page	.164	.029

Put another way, the regression is significant for author Web presence associating positively with Amazon review count. An Amazon author page and Facebook page were significant predictors. The contribution of Goodreads page, Web page, blog and twitter account failed to achieve significance at  $p < .05$ .

*Summary – Popular Sample Regressions*

The following table summarizes the significant predictors for each of the regressions computed on the Popular Sample.

Table 21 - Summary of Significant Predictors, Popular Sample

POPULAR SAMPLE		SIGNIFICANT PREDICTORS					
		Amazon Author Page	Goodreads Page	Web	Blog	Facebook	Twitter
Google	ASIN		X	X		X	
Google	Author/ Title			X		X	
Bing	ASIN			X		X	
Bing	Author/ Title			X		X	
Amazon	Sales					X	
Amazon	Reviews	X				X	

Observations: There is consistency across the regressions computed for each of the dependent variables in the Popular Sample, although interestingly, the significant predictors are different compared to the Random Sample. Both the Bing and Google searches on ASIN were predicted by an author Web site and Facebook page. A Web page and Facebook page were predictive of both the Bing and Google searches on author and title. The Facebook page was predictive for all dependent variables and demonstrated the

highest Betas. Goodreads was significant as a predictor variable for only one of the independent variables. Chapter 5 includes some discussion of the differences between the Popular and Random Samples that may account for these differences.

Of particular note is that all six regression models computed on the Popular Sample were significant and consistent with the hypotheses that author Web presence and participation in social media associates positively with book Web presence, sales and reviews.

### *Ad Hoc Regressions*

This section includes five regressions that were run on selected data treated as independent (predictor) variables that were not included in the original proposed models. These may be of interest for future research, but since they were not part of the predicted models, care should be taken with generalization from the current data sets absent replication and development of explanatory models.

#### 1. Available Print Version as a predictor of Sales, Random Sample

Using the Enter method, a significant model emerged indicating that an available print version of a title contributes to significantly predicting Amazon sales, with  $R^2 = .252$ ,  $F(4,320) = 26.982$ ,  $p < .001$ . Under this scenario, four of the independent variables had significant Betas as follows:

Table 22 - Print Version Betas, Random Sample

Predictor	Standardized Beta	Significance
Print Version	.141	.012
Amazon Page	.133	.009
Goodreads Page	.309	.000
Facebook	.119	.024

Observation: This regression speaks to the issue of whether and to what degree print discoverability may be synergistic with eBook discovery. The results indicate this may be a fruitful area for additional research.

## 2. Facebook Likes, Twitter Tweets and Twitter followers/following as a predictor of Sales, Random Sample

Using the Enter method, a significant model emerged indicating that for titles whose author maintains a twitter account, the number of tweets significantly predicts Amazon sales, with  $R^2 = .078$ ,  $F(3,118) = 3.336$ ,  $p < .022$ . Under this scenario, the number of tweets had a significant Standardized Beta of .243,  $p = .022$ . Neither numbers of accounts following nor number of accounts followed had significant Betas. One issue with tweets more than other factors is that reciprocal causality may make it difficult to separate the independent portion from the dependent portion of the number, since authors tweet, may retweet responses, and then also respond.

For titles whose author maintains a Facebook page, a significant model did not develop for the number of Facebook likes as a predictor of Amazon sales.

Observation: Although number of tweets did result in a significant Beta, these results generally show that the number of tweets or Facebook likes may not be especially predictive of how well authors leverage these sites to improve discoverability.

### 3. Number of books by an Author as a Predictor of Sales, Random Sample

A significant model did not develop for the number of other titles published by an author as a predictor of Amazon sales.

Similarly, a significant model did not develop for a dichotomized variable coded as 0 for an author's only book and 1 if an author had published 1 or more other books as a predictor of Amazon sales.

Observation: this result is somewhat surprising, since some marketing surveys (referenced in earlier chapters) indicate that readers search for books by author, having found an author they like. This result suggests that there may be other factors in play, or that there is a more complex relationship.

### 4. Strength of Predictors as a Function of Fiction/Non-Fiction, Random Sample

A regression was conducted using the predictor variables separately on fiction titles and non-fiction titles.

Using the Enter method, a significant model emerged indicating that for fiction titles, the independent variables significantly predict Google search engine hit counts, with  $R^2 = .241$ ,  $F(6,163) = 8,628$ ,  $p < .001$ . Under this scenario, three of the independent variables had significant Betas as follows:

Table 23 - Fiction Betas, Random Sample

Predictor	Standardized Beta	Significance
Amazon page	.170	.020
Goodreads Page	.253	.003
Facebook	.237	.008

Using the Enter method, a significant model emerged indicating that for non-fiction titles, the independent variables significantly predict Google search engine hit counts, with  $R^2 = .136$ ,  $F(6,148) = 3.868$ ,  $p < .001$ . Under this scenario, only an Amazon page had a significant beta as follows:

Table 24 - Non-Fiction Betas, Random Sample

Predictor	Standardized Beta	Significance
Amazon page	.227	.008

Notably, both Goodreads page and Facebook page drop off as predictive for non-fiction, and the R square value is much lower.

Observation: This result tends to confirm that readers may use different criteria to evaluate fiction and non-fiction works. Goodreads and Facebook continue to emerge as

predictors for fiction title sales, but only the Amazon author page is a significant predictor for non-fiction sales.

#### 5. Total of Web Presence Indicators as a Predictor of Search Engine Counts and Sales, Random Sample

A new scalar variable was created by totaling, for each case, the dichotomous values of each of the predictors, giving for each case a value of 0 through 6 indicating how many of the predictors were used by each author. This value was tested as a predictor using the Google search engine hit count and Amazon sales for the Random Sample.

Using the Enter method, a significant model emerged indicating that the total count of Web presence methods used by an author significantly predicts Google search engine hit counts, with  $R^2 = .164$ ,  $F(1,323) = 63.510$ ,  $p < .001$ . The Standardized Beta for the predictor was  $.405$ ,  $p < .001$ .

Using the Enter method, a significant model emerged indicating that the total count of Web presence methods used by an author significantly predicts Amazon sales, with  $R^2 = .192$ ,  $F(1,320) = 76.016$ ,  $p < .001$ . The Standardized Beta for the predictor was  $.438$ ,  $p < .001$ .

However when the social media count was included with the other predictor variables in a multiple regression using sales as the dependent variable, its beta was no longer significant.

Observation: This result shows that regression is sensitive to the variables included in the model. In this case, the social media count likely correlates with the individual social media dichotomized variables, so other methods would need to be deployed to determine the relationship between number of social media venues used by an author and either discoverability or sales.

### *Correlations*

These correlations were computed to help answer research questions posed in Chapter 1 concerning the potential relationships between certain of the dependent variables.

The various search engine hit counts all measure book Web presence, but in different ways, and the counts for this research come from searches from two different search engines, Bing and Google. The DV counts from the Random Sample were log-transformed and analyzed for correlation by calculating Pearson's  $r$  for each pair of DVs.

Ideally, all of these would correlate positively and strongly. Where correlations fall substantially below 1, it indicates the search engines may not be indexing the same sites, or possibly that precision and recall are less than ideal and vary by type of search or by search engine. Sales were included in the analysis to determine if there was an association between sales and book Web presence, and if so, to what degree.

Table 25 - DV Correlations, Random Sample

			Correlations				
N=326		Bing - ASIN	Google - ASIN	Google - ASIN (Manual)	Bing - Author Title	Google - Author Title	Amazon Sales
Bing - ASIN	Pearson Correlation	1	.738**	.735**	.500**	.278**	.601**
Google - ASIN	Pearson Correlation	.738**	1	.991**	.487**	.246**	.645**
Google - ASIN (Manual)	Pearson Correlation	.735**	.991**	1	.490**	.257**	.639**
Bing - Author Title	Pearson Correlation	.500**	.487**	.490**	1	.502**	.657**
Google - Author Title	Pearson Correlation	.278**	.246**	.257**	.502**	1	.306**
Amazon Sales	Pearson Correlation	.601**	.645**	.639**	.657**	.306**	1
**. Correlation is significant at the 0.01 level (2-tailed).							

The results of the correlations show that all calculated r values are significant at  $p < .01$ , varying in the particular value of r. These show that the correlations between equivalent searches on the two search engines for ASIN are reasonably correlated, that is, the Bing search on ASIN and the Google search on ASIN are measuring the same

construct, to a reasonably high degree, even though Bing returned considerably fewer absolute counts than Google.

A comparison of the automated Google query on ASIN compared to a manual Google search on ASIN is nearly perfectly correlated, indicating internal consistency for Google searches.

Correlation  $r$  values for the Author – Title searches and the Author – Title counts compared to ASIN counts are lower indicating perhaps more noise in precision and recall and/or different indexing algorithms for the quoted search.

With the exception of Google author-title, the correlation of search engine hit count and Amazon sales is fairly high indicating that there is an association between book Web presence and Sales. Further research might shed light on cause and effect.

With regard to the correlation between Google author-title and Amazon sales, this could result if the recall of the Google counts is lower than the other search results. This bears further examination.

### *Analysis of Reviews*

The distribution of the review counts (using the 3-period weighted average smoothing) for the Random Sample and Popular Sample were calculated for a curve estimation. In the Random Sample, 296 (71% of N=419) received no reviews. In order to determine whether the distribution followed a power distribution, the values were transformed by adding 1 to the count of each case and computing the log.<sup>22</sup> These were then placed in rank order and entered into SPSS for curve estimation.

The purpose of the curve estimation was to determine which class of transforms could be used to correlate review count with search engine hit count. The results of this correlation directly address the hypotheses that these constructs would associate positively as representing a measure of book Web presence, or discoverability.

The results of the curve fitting show that consumer review counts closely follow a power law distribution of the form

#### Equation 6 - Sales Rank as a Function of Review Count

$$\text{rank} = a * \text{count}^n$$

where  $a=112.9$ ,  $n= -.849$  with  $R^2=.894$ ,  $F(1,414) = 3500.915$ ,  $p<.001$  as illustrated on the following graph.

---

<sup>22</sup> This is a standard transformation technique when data contains zero values, for which the log is undefined.

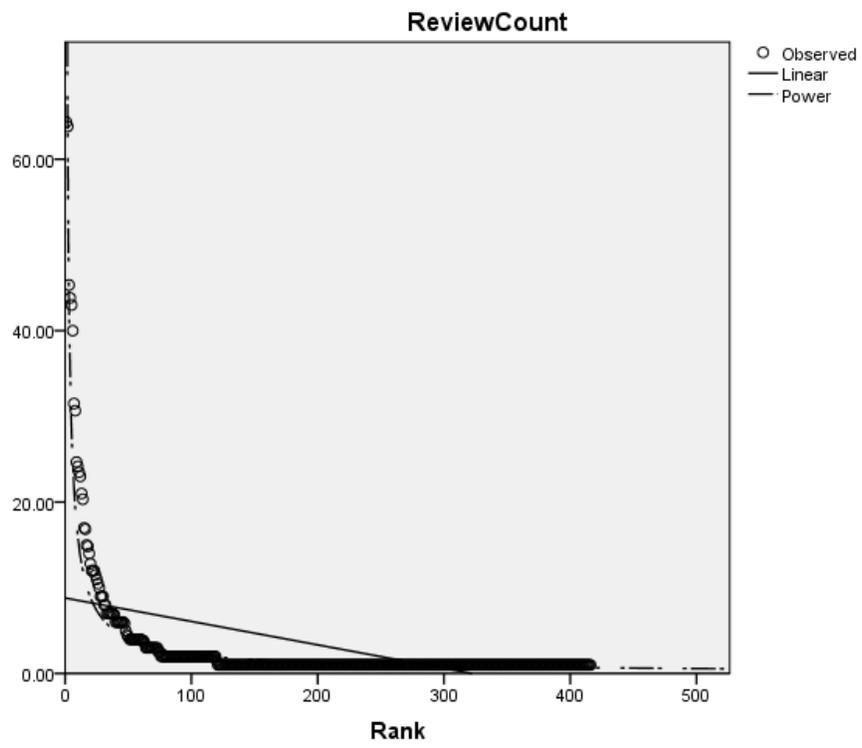


Figure 4 - Plot, Rank vs Review, Random Sample

Similarly, the review counts from the Popular Sample significantly fit a power curve with  $a=179863.994$ ,  $n=-1.792$ ,  $R^2 = .901$ ,  $F(1,181) = 1645.302$ ,  $p < .001$  as illustrated on the following graph.

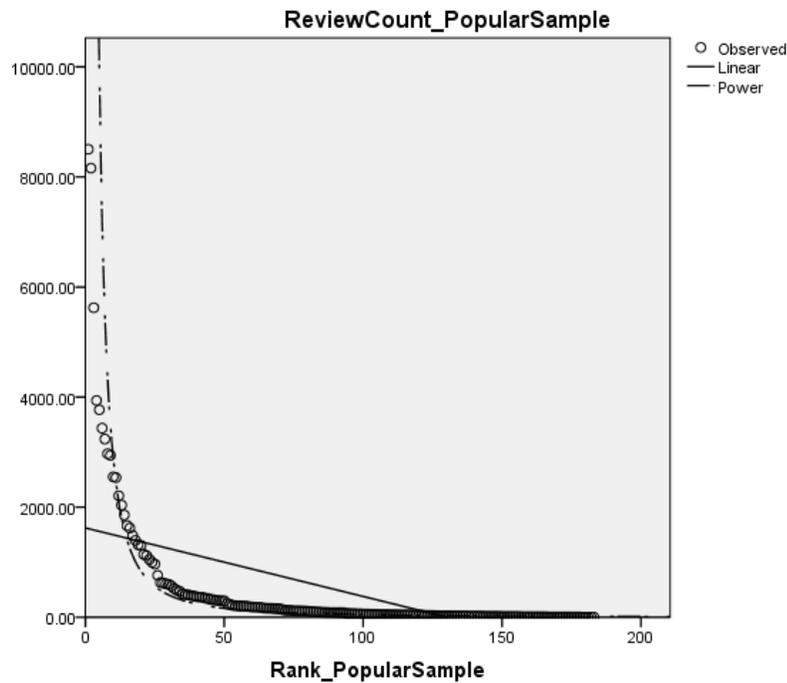


Figure 5 - Plot, Rank vs. Review, Popular Sample

A Pearson product-moment correlation coefficient was computed of the log transformed variables to assess the relationship between review counts and the Google search engine hit count on ASIN. For the Random Sample titles selected for the regressions, there was a moderate correlation of  $r = .606$ ,  $n = 326$ ,  $p < .001$ .

For the Popular Sample titles selected for the regressions there was a high correlation of  $r = .860$ ,  $n = 179$ ,  $p < .001$ .

*Research Questions and Hypotheses – Phase II*

RQ3. To what extent are eBook authors and publishers establishing Web presence, and is author Web presence differentiated by sub-category of book?

The data show that authors establish Web presence and participate in social media to varying degrees. The data show that successful authors, as represented by the Popular Sample, have established Web presence at a higher rate than the typical author as represented by the Random Sample. The results are less clear when the Random and Popular Samples are categorized into subgroups. For example, although the Random Sample has a higher percentage of self-published titles, there does not appear to be a significant difference between the self-published and mainstream-published titles with respect to social media participation.

RQ4. To what extent does author Web presence account for search engine page hits and sales? Does its effectiveness vary by sample sub-category, such as genre or self-published vs. mainstream published?

The regressions show that author Web presence associates positively with and predicts a portion of the variance observed in counts of search engine page hits and sales. All regressions models for author Web presence were significant, and there was consistency in the emergence of certain predictors as significant depending on whether the regression is computed on the Random Sample or the Popular Sample.

On the other hand, there was little differentiation by sub-categories; that is, while significant models developed for certain sub-categories as evidenced by the ad hoc regressions, the standardized Betas did not differ appreciably from the undifferentiated results. The single exception was that presence of a Goodreads page was not as much of a predictor for non-fiction titles compared to fiction titles.

RQ5. Is there a relationship between sales, measured by sales rank, and book Web presence, measured by search engine links returned?

There is a moderately high positive correlation between sales and all book Web presence dependent variables. This may be result of sales driving Web presence, Web presence driving sales, or some combination of both.

H1: Author Web presence associates positively with eBook Web presence, as measured using search engine result counts with high precision queries.

H1 is confirmed. All regression models testing the association between author Web presence and book Web presence were positive and significant. The effect size ranged from low to moderate. Not all predictor variables were significant, and the significant predictor variables were different for the Random and Popular Samples.

H2: Author Web presence associates positively with eBook sales, as measured using Amazon Sales Rank

H2 is confirmed. All regression models testing the association between author Web presence and sales were positive and significant. The effect size

ranged from low to moderate. Not all predictor variables were significant, and the significant predictor variables were different for the Random and Popular Samples.

As noted in Chapter 1, some kinds of social messages are not individually accounted in search engine estimates of message counts. These include, for example, consumer review counts on sites such as retailers Amazon and Barnes and Noble, and social book sites such as Goodreads. Consumer reviews are an important component of social gatekeeping message creation and this raises certain questions and a hypothesis:

RQ6. What is the relationship, if any, between book Web presence and consumer review count?

There is a moderate to high degree of association between book Web presence and consumer review counts indicating that consumer review counts may serve as an indicator of book Web presence.

RQ7. Do rates of diffusion of information on the Web and rates of numbers of reviews correspond over time?

The correlation between Web presence and review counts is markedly higher for the Popular Sample compared with the Random Sample. The Popular Sample titles have been on the market for some time compared to the Random Sample only on the market for 15 weeks. This may indicate that the association between search engine hit count and review count may increase (converge) over time. This would take further research to confirm.

H3: eBook Web presence associates positively with consumer review counts.

Analysis of the rank distributions confirms that both Web presence as measured by hit counts and consumer review counts follow a power law curve. The correlations show that eBook Web presence as measured by hit counts and consumer review counts associate positively to a moderate to high degree.

### **Phase III - Title Review**

#### *Purpose*

The purposes of the third phase of data analysis were to investigate selected authors, publishers and titles for additional insight into the use of social media to improve discoverability and sales, to review the various methods of social network engagement to see whether additional independent variables might surface that could impact the interpretation of the regressions, and to look for evidence of behaviors or technologies suitable for further research on the proposed serendipitous tie construct.

#### *Research Questions – Phase III*

An open-ended review of author Web presence and use of social media in a selected (non-random) subset of the Random and Popular Samples beyond those identified as independent variables in the first two phases of research was conducted in order to address the final two research questions:

RQ8. What additional insight can a more thorough examination of the samples provide that informs interpretation of the results of the descriptive and inferential portions of the analysis?

RQ9. What does a more thorough examination of the samples suggest for future research?

*Review of Selected Titles from the Random Sample*

B007QQDTSQ *The Bread Man: A Kindness Never Forgotten* by Will Bevis, self-published. This is a single short story of approximately 23 pages, one of several shorts in the Random Sample. This author has a number of shorts available in Kindle format he calls e-stories, many under 10 pages, and at one point advertised on his Amazon author page that he was providing free downloads for all his older titles, about 50 in all. The author maintains an extended biography on an Amazon author page. His Facebook page was nearly unused with only 6 likes. He was active on Twitter, but at the time his twitter account was reviewed, neither it nor his Website appeared to reference the title under review. He appears to use Amazon exclusively and his Website currently provides links only to his Amazon author page. He did not appear on Goodreads. His closing sales rank for this study was 685,914 indicating only a very small number of sales.

Comments: a number of authors are experimenting with short story fiction, releasing each story as a low priced eBook. This is something that would likely not be economically feasible in print and may thus represent a new class of literature packaging exclusive to the digital market. However, few of the short story singles attracted more than a handful of readers. Amazon provides only spotty tracking of length, so it was not possible to quantify this category using the data collected, but it bears further research to describe this market segment.

B007QI2FAC *Blood Brothers Book Two: Warrior's Journey* by Sadie and Sophie Cuffe, hybrid published. Desert Breeze Publishing uses a hybrid publishing model that does not pay advances but does pay royalties. Strictly speaking, it does not appear to be

author subsidized. It offers some promotions, but appears to expect the author to do most of the marketing and now requires accepted authors to maintain a Website. At the time the title was reviewed, the authors (sisters) did not maintain a Web page nor an Amazon page or twitter account, but did appear on a Goodreads author page. Their Facebook page had acquired 40 likes. Closing sales rank was 610,330 indicating only a very small number of sales.

Comments: This title is representative of a handful of titles in the Random Sample coming from what appear to be hybrid publishing models featuring publishers that organize around a theme (in this case, non-erotic romance) but who do not pay advances and who expect authors to engage in self-marketing. The authors did little in the way of social media outreach in the first several weeks.

B007R8BFHU The Naughty Pleasures Bundle by Abbie Cole, self-published. This title ended the data collection with the best sales rank of any self-published book in the Random Sample, 7,300, indicating brisk sales for at least several weeks prior. It is a bundle of three prior titles each with 3 erotic short stories for a total of nine short stories. The author made little use of social media, with only 1 like on Facebook, and no Web, Twitter, or Blog. There is an author page on Goodreads but with little information other than her name, however. The title was still selling reasonably well a year following publication.

Comments: This may be an example of a title doing well simply by virtue of its classification as erotica with high visibility in that category on Amazon. Social media outreach was minimal and there is little else to explain its popularity.

B005GSZZ2Y Sweet Addiction by Maya Banks, published by Berkley/Penguin.

This title ended the data collection with the best sales rank of any traditionally published book in the Random Sample, at 927 indicating very brisk sales for at least several weeks prior. It is erotic fiction in the BDSM niche, possibly benefitting from the popularity of the best-selling 50 Shades of Gray trilogy, also in the BDSM niche erotica genre. The author maintained an Amazon author page with little information but a link to her Twitter account with 16,991 tweets (up to 20,775 a year after publication), a very active Facebook account with 7846 likes (up to 18,174 a year after publication), a Web page and a Goodreads author page. While the Facebook page is primarily devoted to wall posts about her books and writing, the Twitter account posts consist of several daily tweets on a variety of subject including but not limited to her books and writing. The Website is extensive with links to the Twitter and Facebook accounts, and a link to an age-restricted Yahoo group discussion site. She maintains an active Goodreads author page that links to her website, Facebook and Twitter page.

Comments: The author is highly visible in social media, had built up considerable social capital either before or within a very short time after publication of this title and actively promotes her work on the social media sites. Several titles categorized as erotica did relatively well in sales, although many, especially the shorts of a few pages or less, fared more poorly.

B007EFHQHE The Mother Road by Jennifer ALee, published by Abingdon Press (an imprint of the United Methodist Publishing House). This title is one of two from the Random Sample that also appeared in the Popular Sample indicating a large

number of downloads immediately following publication. The book is classified as Religion and Spirituality – Fiction. The author is active in all of the social media examined. The Facebook page was self-classified as a Readers' Group page with 203 likes but only a handful of posts since joining Facebook in 2011. Her Twitter account registered 1771 tweets at time of initial review (up to 1876 a year later), almost all about her books and writing. The Website redirects to a combination Web and blog hosted on Blogspot. The Website links to Facebook and Twitter, provides purchase links on Amazon and two Christian book sales sites, hosts a calendar of book signings and book club appearances, and offers an emailed newsletter by subscription. Her Goodreads page links to her Website. Her closing sales rank was 60,156 indicating modest sales.

Comment: A number of titles in the Christian genre – both fiction and non-fiction – were included in the Random Sample. The social media outreach for this title seems modestly developed but there apparently is a loyal fan based for these kinds of works. This is one of only a few sites where book signings and book club appearances were explicitly referenced. The extent of the mailing list was unknown, but may have contributed to the nearly immediate download response, along with the possibility of outreach by the publisher, an imprint of the United Methodist Church.

B005FLODDE *The Woman Who Wasn't There* by Robin Gaby Fisher and Angelo J. Guglielmo, Jr., published by Touchstone / Simon and Schuster. This title is one of two from the Random Sample that also appeared on the Popular Sample paid list indicating a large number of sales immediately following publication, with an initial sales rank of 66. Publication of the eBook coincided with a hard cover print edition. It is

biographical non-fiction and tells the story of a woman who claimed escape from the 9/11 event, her life, and the fraud and deception that followed. Nearly a dozen editorial reviews or quotes from noted sources are provided on the Amazon page. The first author, Fisher, did not register on Amazon but the second author, listed as coauthor, did. Fisher's biography notes that she is a two-time Pulitzer finalist in the feature writing category and had two New York Times bestsellers prior to release of this title. The Facebook page lists both authors but has few likes and only a handful of posts. It was set up just days before the release of the book. The Twitter account is similarly modest, in Fisher's name only, with sporadic posts coinciding primarily with book releases or events. The current Website is simple and contains links to her books and an email address but no other linkage to social media accounts. The current Goodreads author page provides a short biography and a link to her Website, but does not appear to have been updated prior to publication of this title. The book was still selling well a year later with a sales rank of 16,917.

Comments: This is representative of a number of titles in both the Random and Popular Samples of mainstream mid-list authors who invest relatively little effort on social media but nevertheless do well in sales, presumably on the basis of reputation and positioning directly on the sales sites, and the strength of mainstream media and trade reviews and promotions.

B007RHLVPW Tales of Tomorrow Girl by Robert Szeles, and

B007RJFY5I The Never Men by Robert Szeles, self-published. Tales of Tomorrow Girl is a collection of six short stories around a science fiction theme; The

Never Men is one of those six short stories. Each of the individual stories is about 20 pages in length and sold for \$0.99; the collection sold for \$2.99. Investigation shows that all six of the short stories were individually released on the same day, although only one showed up in the Random Sample. The author maintained an active Amazon author page with links to a Twitter feed. The Facebook page was started in 2011 and contains few posts. None of the posts reference either of these two titles and the account was basically inactive between March 2012 and March 2013. It does, however, link to his Amazon and Smashwords author pages. The Twitter account is slightly more active but contains only a small number of tweets coinciding with the release of these titles. The Website primarily promotes a non-science-fiction novel released before these titles. The site contains news, links to Facebook, Twitter and his blog site. It also contains a brief biography and calendar, updated to mark the release of these titles but with no updates since. The Goodreads author page links to the Website and Twitter account and includes a biography. Reader response was poor. Final sales rank for the collection was 780,117; for the short, 878,730.

Comments: This is one of several examples in the Random Sample of authors dividing up and combining releases of short stories in collection and single releases. The author seems primarily interested in promoting his novel-length book and so there is relatively little about these titles in his social outreach. This is one of a handful of Amazon author pages that links to a professionally produced video book trailer, albeit for his full length novel and not these titles.

B007MEUTS0 *The Ethics of Business: A Zondervan Digital Short* by Scott Rae, published by Zondervan, an imprint of Harper Collins. This digital short, at an estimated 36 pages, is condensed from the textbook *Moral Choices* by Scott Rae and was originally entered in the Amazon database with Zondervan as the author as well as the publisher. It is apparently intended to serve as a discussion guide, although the description is lacking. No social media engagement was found. It ended up with a final sales rank of 552,575 indicating few sales.

Comments: This title appear to be an abbreviated version of previously released content. Repurposing of content such as extended media coverage on a subject originally published in magazines or newspapers and repurposed as an eBook, or eBooks such as this one condensed from a title of greater length are a relatively new trend, which along with the digital short stories may represent a new category of reading material finding a market on digital devices.

B007O04CZI *Preparing for the AP Biology Examination: Fast Track To a 5, 1st Edition* by Robert Doltar, published by Delmar Learning. This title was pulled or unpublished on Amazon almost immediately, but was republished the following week in Kindle and print formats with Cenage updated as the publisher of record, and it was republished a third time with a 2013 publication date and copyright. Goodreads shows a version first published in January of 2012. Delmar, which is owned by or is a subsidiary of Cenage, specializes in writing textbooks and curriculum to spec for educational institutions, state and local government and the private sector. Because the original

version as selected was unpublished, it was eliminated from the Random Sample, and was not further tracked.

Comments. This title is representative of a number of titles from the Random Sample that were unpublished or removed over the course of data collection. Some of them, like this title, were republished with new ASINs under a different publisher, with or without revised copyright dates. Others were republished with no apparent changes, and some simply disappeared. These examples show the ease with which modern digital publishing enables updates, revisions and changes, which can also cause problems for those concerned with bibliographic issues such as librarians.

B007FFKV9S *Cast of Characters* by Lou Aronica, self-published with a twist, under the imprint The Fiction Studio. This title is an anthology of short stories written by multiple authors and co-written by agent/author/publisher Lou Aronica, owner of The Fiction Studio. Among the company's other titles are several books written by Aronica, several listing Aronica as co-author and several listing either Aronica or Novelists Inc. as author or publisher. According to his Facebook bio, he has considerable experience in the publishing industry including senior positions at Bantam, Avon and other traditional publishers; he established The Fiction Studio in 2003 as "a creative development company with a publishing imprint for new authors." Some of the titles on the company web page are his or his in co-authorship or editorship with other individuals. There are also books written solely by other authors. For a few examined, the book preview shows The Fiction Studio listed as publisher but the author as copyright holder. One of the Popular Sample titles also comes from The Fiction Studio (B005SZ0W14). No contract

details are listed on the Web site but there is an invitation to email him for details on their “publishing program.” Aronica’s Facebook page is actually auto-generated from his Wikipedia entry. His twitter account is modest and mostly promotes books available on the Website. His Goodreads page links back to the Website and lists 23 separate works by Aronica alone. This particular title sold modestly well with a final sales rank of 69,530 at close of data collection.

Comments: Slowdowns, shakeups and consolidation in the publishing industry have led several former mainstream agents and publishers to enter the self-publishing or niche publishing market, either as publishers, writers, agents, consultants or pundits. Perhaps the best known of these is Mike Shatzkin, former agent and founder of The Idea Logical Company specializing in consulting and writing about trends in digital publishing. Aronica appears to be one of those former publishing executives who are developing new business models that straddle self-publishing and traditional publishing. This title is one of a small number in the Random Sample that appears to follow a small hybrid publishing cooperative business model.

B007RV7K5S Five Hands Diet by Emily Callowdic, self-published. This title runs 3 pages total including, presumably, the title page and is priced at \$4.95. According to the Amazon description, the book details a particular diet approach with sample foods and recipes, but according to a contributed consumer review, the contents include only a small list of foods with calorie counts. There is no social media outreach noted. The title nevertheless has sold at least a few copies with a final sales rank of 179,492 at close of data collection.

Comments: A number of very short works appear in the Random Sample, numbering 10 pages or less. There have been unsubstantiated rumors in blogs published in spring 2013 that Amazon might start rejecting very short eBooks, under 2,500 words, but that could not be confirmed as this is written. Many of the books in the erotica genre of the Random Sample are shorts of only a few pages. Most in the erotica genre use social media more extensively than this particular title.

B007R3Y1EO Drive Business Growth with the Telephone by Valerie Schlitt, self-published. Schlitt runs a telemarketing firm and the Website promotes the services of the firm without mention of the book, as do the Twitter account, little used, and the Facebook page, with more posts but few likes and comments from third persons. No book-related social media use was located. The sales rank of 964,418 indicates only a very small number of sales.

Comments: This book is one of a handful in the Random Sample that links to a business venture, often consulting. The purpose of the eBook may simply be to add awareness and recognition to the business without ever expecting significant sales. It may also be that a version is available for clients free through some other venue, although that could not be determined in this case. Use of social media seems low for someone running a telemarketing consulting business.

B007R7FAU4 The Artist Charge (A Variety of Passion) by James Baumann, self-published. This is one of 2 single-page “e-poems” found in the Random Sample. No social media outreach was located, and lack of a final sales rank indicates no sales.

B007RS202E The Gods of the Lodge by Reginald Haupt Jr., published by RiverCrest Publishing, originally published in 1990 by Victory Publishing Company. RiverCrest is a niche publisher specializing in conspiracy oriented non-fiction, and this title details alleged conspiracies in the Masonic lodge. The author is not in evidence nor could any social media be located related to him or the title. Some promotional material was located related to the RiverCrest author Texe Marrs who wrote a new forward for the book.

Comments: Some titles in the Random Sample are reprints of recently published books not in the reprint jungle category. This particular title appears to have no involvement from the original author and may be something acquired by the publisher from a third party.

B007R0HH8O Complete Defense to Queen Pawn Openings by Eric Schiller, published by Cardoza Publishing. This title was originally released in paperback format by Cardoza Publishing in 1998. The eBook is a re-release by the original publisher, but without a re-release of the print edition. The author maintains an active Facebook and Web page devoted to chess where his 1998 print book is listed, but the Web site has not been updated with the eBook version. There is no Twitter account, and a Goodreads page exists but has not been claimed.

Comment: This is one of a handful of books re-published in eBook format by the original publisher. In this case, the author does not appear to have promoted or been involved with the new eBook version.

B007QR52C6 *Tess and the Star Traveller* by Jane McKay, self-published.

Originally published by Fremantle Arts Center Press, a non-profit publisher specializing in Western Australian people and culture, this title is now self-published by the author. Social media on first review included only the Amazon and Goodreads pages plus a Website. The website currently lacks any reference to the book, instead featuring her paintings and other artwork, as does a new twitter account. The Goodreads page contains no author information and lists only the one title. No sales were recorded during the data collection period, and a year later, there is still no sales rank.

Comments: This title represents one of only a couple of examples in the Random Sample of self-publication after initial publication by a mainstream publisher who turns publishing rights back to the author (so-called rights reverted to author). For another in the Popular Sample demonstrating rights reverted to author, see B003TU2JJ8. Many more examples of this kind were found in Bradley, Fulton and Helm's 2008 snapshot of self-published titles, so the lack of more examples in the data was a surprise. A more in-depth examination of bibliographic records might find more cases in the Random Sample. This title is also representative of a handful of authors in the Random Sample who used their social media outreach for pursuits other than promotion of their eBooks.

B007QP4D6Y *Resonance & Vengeance* by A.J. Scudiere, self-published.

Originally featuring only a Twitter account and Website for social media outreach, Scudiere is now one of several authors in the Random Sample experimenting with social media other than those studied here as IVs. The Website currently includes links to an Android mobile phone app that rebroadcasts a twitter feed and audio podcast (not present

during the initial 2012 review), in addition to links to Facebook, Twitter and Goodreads, and a new blog, also not present in 2012. The book is actually a compilation of two books previously released in 2008 and 2010. At the close of data collection for this study, the title ended up with a sales rank of 880,540 indicating very few sales.

Comment: Several authors have increased social media outreach since the titles were first reviewed in 2012 after the 15<sup>th</sup> week of publication, but only a handful have gone beyond Facebook, Goodreads and Twitter for social media outreach.

B005K0HDGE 11/22/63 by Stephen King, published by Scribner, Popular Sample. King's website and all other social media outreach is managed by a company under King's ownership and direction. King does not personally manage the pages and had an FAQ at the time of review stating, "Stephen does not have any social network pages, and has no plan to set any up." However, there is a Facebook page maintained on his behalf with 490,000 likes, and a Goodreads page with biographical information entered by Goodreads volunteer editor/librarians.

Comments: Several authors in the Popular Sample rely on their publishers or agents to maintain their social outreach.

B004A90BXS Heaven is for Real: A Little Boy's Astounding Story of His Trip to Heaven and Back by Todd Burpo, published by Thomas Nelson. Burpo is a minister who wrote this account of his son's near death experience. Burpo does not appear to aspire to a writing career although some other books relating to this title have been released (a study guide, children's versions, and a German translation). All of the social outreach is

ted to the book title rather than the author, for example the Facebook page and twitter pages are indexed under /heavenisforreal and the website is <http://heavenisforreal.net>.

Comments: a few cases in the Popular Sample were found where authors had created social media sites identified by book title in addition to maintaining personal social outreach pages. This title appeared to benefit from wide mass media exposure and remained a top seller throughout the data collection period ending up with a sales rank of 148.

*Briefly Noted from the Popular Sample*

B006POB270 by Elsie - Adventures of an Arizona Schoolteacher 1913-1916 by Barbara Anne Waite. Her Website is one of only a few with links to her page on the social site LinkedIn.

B00514OWKY Whispers in the Sand by Barbara Erskine. The Website links to Stumbleupon, which allows guests to share on that social network.

B005BU9JK6 Gateway to Heaven by Beth Kery. The Website provides share links to social networks Tumblr, Pinterest and Google+. Note that these are not sites maintained by the author but rather facilitate a guest sharing the Website content to their own accounts on these sites.

B003SE7J6I Beneath a Buried House: A Detective Elliot Mystery by Bob Avey. Site embeds Goodreads reader reviews, provides an audio excerpt from one of his books, Maintains and links to a personal Pinterest page, also links to Google+ and Linked-in.

B0050KTQ0K *Divorced, Desperate and Delicious* by Christie Craig. The author also guest blogs on a fiction blog site and maintains a writing workshop website.

B005BUG6TI *Betrayal: A Novel* by Danielle Steel. The Website provides a social share widget that permits guests to share on virtually any social network. There is also a link to a Pinterest page maintained by the author or her management company. The Facebook page has nearly a million likes.

B007J4T2G8 *Fifty Shades of Grey* by E L James. The Website links to the author's Youtube channel where she has set up playlists featuring potential or planned soundtrack titles possibly related to upcoming movie versions of the books.

B00492CK1M *Spying in High Heels* by Gemma Halliday. The Website links to the author's Youtube channel which includes video book trailers for some of her works, in addition to a collection of unrelated "favorited" videos.

B000QCS8TW *A Game of Thrones: A Song of Ice and Fire: Book One* by George R.R. Martin. The Website is extensive and professionally maintained with links to merchandise, the HBO series, social media sites, news, appearances and other pages or separate Websites related to the author's work. The site is almost certainly maintained by a management company.

B004ULORYU *Killing Lincoln: The Shocking Assassination that Changed America Forever* by Martin Dugard. The author's Website now links to an Instagram account, but the account has no posts or pictures and only 2 followers.

B007QVABM8 *The Hunger Games #2 Catching Fire* by Suzanne Collins. The Website is surprisingly sparse for a front-list bestselling author. It has a homemade look and contains only 4 pages of information about her books but nothing on the movie versions or merchandise tie-ins. Her Facebook page was (and still is) auto-generated from Wikipedia (one of several, in both the Random and Popular Samples) with no personal involvement, but it still had generated 109,000 likes when reviewed during data collection (currently up to 186,000). No Twitter account was found during data collection. A Twitter account is currently active, but with only 35 tweets dating back to January 2013 and only 322 followers.

B007MTR4AQ *Ball Four* by Jim Bouton, published by RosettaBooks. RosettaBooks specializes in republishing back-list and mid-list titles in eBook format and has acquired rights to republish several well-known authors such as Huxley and Vonnegut. This title was originally published by Bulldog in 2001. This is an example among many of older print titles republished as eBooks.

B005T634QC *Cloak* by James Gough, published by WiDo Publishing. The publisher is a small niche company serving primarily young adult fiction and fantasy. The publisher follows the mainstream model, but makes the following statement on the guidelines page for submissions: “We expect our contracted authors to come prepared with a marketing plan outlining the specific ways they intend to promote their book. Even a work of fiction needs a platform that can be maximized for promotion. Social media is key in book promotion and visibility, and we expect each author to be active in social media before and after the book’s release. A strong social media presence will be in your

favor as we consider your manuscript for publication.” (Retrieved 4/15/2013 from <http://widopublishing.com/submissions/>).

B007JC2942 Playbook 2012: Inside the Circus--Romney, Santorum and the GOP Race by Evan Thomas, published by Random House. This title is representative of a small number of titles in the Popular Sample released by a Major publisher only in eBook and audio format.

B006NFB5BY Solo by Sarah Schofield, self-published as M.O.I. Publishing. This title represents a small number of books coming from what appear to be cooperative self-publishing companies. Two authors participated at the time of the initial review, and a third has since been added. The print versions came through Createspace indicating the self-published origins, but the authors have set up a web site featuring the publishing firm, presumably for purposes of marketing. The author also maintains an individual Facebook page and an individual Web page with the first two chapters of the book uploaded to the file sharing site SCRIBD.com.

### *Summary*

The review of selected titles with respect to publishing models and social media outreach suggests several potential trends and indicators that both inform the current inquiry and lay the groundwork for future quantitative analysis.

It was clear during examination of individual titles and authors that there is considerable variation in how social media is used to promote books and book discovery. For the quantitative portion of this research, the social media independent variables were

dichotomized as 1/0 (yes/no) with no factors taken into account about how effectively social media was used. In fact, social media use even for those with equivalent dichotomized profiles may have made very different use of social media, not only in how much and how often they posted, but in what authors posted about, how they invited reader interaction and engagement and how well they managed to coordinate their outreach efforts (for example, by linking each of the social accounts to each other so that guests might easily navigate from one site to another).

Relatively few authors pursued social media other than Facebook, Twitter and Goodreads, and even many of those who maintain presence on these sites failed to link to them from their Website or from each other. A few authors are now experimenting with Pinterest (image oriented), Stumbleupon (a discovery-oriented search engine), Youtube (videos) and LinkedIn (primarily used for business contacts). Links to Librarything (a book network similar in functionality to Goodreads), the author's Amazon author page, and maintained presence on other social sites was and still is infrequent.

In comparing primarily Websites visited during the initial review versus those same sites a year later, there did appear to be an increase in the use of social site sharing widgets, which allow guests an easy way to update their own social network pages. That is, the user can click on the widget corresponding to a social site on which they have an account, which then logs them in and creates a post such as a "like" or status update linking back to the author's page. These can be programmed to provide information back to the Website host as well, and some companies provide widgets with analytics that can inform the host of what sharing has taken place (e.g. addthis.com).

While a few best-selling authors take a hands-off approach to social media marketing and let management companies, agents or publishers maintain the social media outreach (or eschew it altogether), most authors in the Popular Sample appear to be active participants. There is a marked difference in the rate of social media participation, in terms of whether an author establishes presence on the different social sites, between the Random Sample and the Popular Sample. Part of this may be effect as well as cause, and it also may be a result of shifting strategies even among mainstream model publishers. At least a handful of mainstream model publishers reference an author's responsibility to maintain social media outreach on their submission guidelines pages, and at least one publisher would only consider submissions from authors with an established social presence.

## **CHAPTER 5 - DISCUSSION**

### **Introduction**

This final chapter first summarizes and recaps the purpose of the research and key points of the theoretical framework developed that guided the research agenda and methodology. The results of the research are then summarized, conclusions are drawn and the potential significance of the findings are discussed. Finally, the implications for future research and recommendations for action are presented.

## **Summary of the Purposes of the Research, Theoretical Framework and Methodology**

This section provides a synopsis of key points made in Chapters 1, 2 and 3.

Technology and the nearly universal adoption of computer mediated communication channels such as the Web have reduced or eliminated both the costs and the commercial gatekeeping barriers to the publication and dissemination of books, both in traditional print format and in digital formats. Mainstream book publishing – once limited to companies with sufficient capital to pay author advances, manage expensive printing presses, build warehouses, control distribution networks and negotiate with retail outlets for marketing and visibility – has become democratized. Today, any aspiring author with a connection to digital networks and a few inexpensive tools can create, format, publish and sell cheaply (or even give away) books to the public.

The result has been an explosion of new titles from traditional mainstream publishers, niche-market and specialty publishers, and self-publishing authors. By 2012, the number of self-published titles entering the market each year reached and even exceeded the number of traditionally published titles, and self-published titles occasionally achieve best-seller status. The full extent of this glut of titles is unknown. Until recently, virtually all books available in the commercial/retail space were issued ISBN numbers that could be catalogued and tracked by the issuing agency. Today, many books and eBooks sold online never acquire an ISBN number and there is no central tracking authority. The result is that over the last few years, millions of new titles have

been released in multiple print and digital formats that can be found in a multiplicity of outlets including online store fronts and independent outlets, both physical and virtual.

This hyper-abundance of titles in a heterogeneous marketplace gives readers unprecedented choice, accompanied by the newly emerging problem of discovery: how do readers manage to find books they want and how do authors make their books discoverable so that they stand out among readers?

As books, and especially books in digital format, are increasingly discovered and retrieved in virtual storefronts hosted on the Web and on other digital networks, information about books – the book metadata – moves online as well. Authors and publishers both must find ways to reach readers through these channels to inform them about books they may want to read. Traditional publishers with sufficient capital for advertising and marketing still may avail themselves of the mass media. For many authors and small publishers, however, who lack sufficient capital to invest in mass media campaigns, computer mediated communication channels are among the means through which readers become aware of books that may interest them. Marketing research suggests that readers find books in a few different ways, such as browsing, recommendations from friends, mass media, and social networks. The lens through which this milieu is examined in this research is *Social Gatekeeping*, an extension of traditional gatekeeping theory that provides a framework for understanding how author participation in social networks and other computer mediated communication channels initiates a flow of the diffusion of information over the web and other computer mediated

communication channels and through individuals and social networks to potential readers.

The traditional publishing chain is the sequence of gatekeeping decisions through which a book and its metadata progresses from author to agent to publisher to distributor and finally, retailer. At each step, a gatekeeper makes decisions about aspects of the book and adds value to it. In the traditional publishing model of the 20<sup>th</sup> century, the publisher stands as the primary gatekeeper directing the flows of the book and information about the book (the book metadata) as it works its way to the retail channel. While the book is working its way through editing, design, printing and distribution channels, information about the book is made available through mass media advertising and marketing and through a mass media network of reviewers and critics. The reading public either finds it directly (a reader finds the review or sees the ad, or finds the book in a library or on a promotional table inside a bookstore) or indirectly from a person of influence (a friend or respected acquaintance) who has learned of the particulars from the mass media. Then, word of mouth spreads the information through social networks

Traditional publishers have always relied on social networks and word of mouth diffusion to reach readers. For mainstream publishers of the 20<sup>th</sup> century, the diffusion of information began primarily through mass media, with a smattering of direct-to-consumer marketing and direct marketing to persons of influence such as book club leaders. However in the 21<sup>st</sup> century, computer mediated communication facilitates direct discovery of books through online venues that bypass mass media gatekeepers and provide new mechanisms for the flow of book metadata. This includes discovery at point

of sale, such as online markets, and discovery through computer mediated venues such as Web sites, blogs, social network fan pages and other online social information constructs.

While rejected authors often think otherwise, gatekeeping is not necessarily bad. Gatekeeping serves to filter out poor quality while promoting the best work by the best authors, made better by the services of experts in the publishing chain such as editors and artists. There is also a sense of gatekeeping as information intermediary where gatekeepers serve the role of finding information for those seeking it. Librarians and other information professionals model gatekeeping in the sense of information filtered in.

An expanded view of the publishing chain as it exists at the beginning of the 21<sup>st</sup> century shows that technology and access to computer mediated communication has greatly reduced costs and barriers to reaching readers and readers' social networks directly, and authors and publishers no longer are constrained by the traditional publishing and mass media gatekeepers. Mainstream traditional gatekeepers still exist and many readers trust them to provide high quality reading experiences by virtue of the gatekeeping process but now, technology, computer mediated communication channels and social networks have made it possible for individuals to act as gatekeepers to their friends and acquaintances as well as to the Web browsing public.

When a person learns about a book (or other product or service), that person may choose to let friends and acquaintances know about it, or conversely, the individual may choose not to pass the information along. This has always been the case, even in face to face communication, but the technologies of the Web and online social networks have made it easy both to publish the information and to make it more broadly available to

friends and acquaintances and even to others not directly connected to one's social network.

This is what is called, in the context of this research and framework, *Social Gatekeeping*, defined to be the process of finding, selecting, filtering and shaping information about a product, service or idea and making it available as a message accessible in a social communication channel. Further, social gatekeeping may be performed not only by individuals but by processes that filter, select and recommend based on social data.

Information diffusion through face to face social networks has traditionally been described in terms of the strength of ties between members of a network, with close friends representing strong ties and more distant acquaintances representing weak ties. It has been known for some time that new information most often comes from an individual's weak ties, and this has been shown experimentally to be the case. On the Web and within social networks, new information may come through browsing activities from sources that are not explicitly linked to an individual's social network, even weakly. Or, information may come to an individual from a process or application that leverages social information. For purposes of this research, this is a serendipitous tie, an incidental, chance or accidental interpersonal relationship event that may occur between people not otherwise socially connected, by means of which information may be passed and communicated from one individual, and potentially one social network, to another individual and social network.

For the author or publisher, social gatekeeping may be a strategy that supplements – and in some cases supplants altogether – the role of mass media in making information about a book visible and that triggers the diffusion of information about the book through social networks and generally through the Web. The diffusion of information about a book can be measured a few different ways. Search engines can be used to perform specific queries about books and a count of the returns will reflect book Web presence if the query has high recall and precision. Sales can be directly compared if known, or inferred from estimates derived from sales rank. Direct evidence of social gatekeeping may be reflected in counts of consumer reviews. Although it may not be possible to count every instance of relevant web pages, sales or reviews, sampling them consistently and without bias provides a means of comparing titles and estimating the extent to which certain factors might predict greater diffusion.

Authors have many potential pathways to enhance discoverability for potential readers. Authors working with traditional publishers or engaging in self-promotion and marketing may still leverage mass media. Authors may also work individually or in concert with publishers to arrange book tours, meet with book clubs and manage direct marketing. Self-publishing authors and publishers can also experiment with targeted pricing and pricing promotions on sites such as Amazon. The principle organizing research questions posed by this framework are: how and to what extent do authors connect to readers through social media, and what is the extent to which such use increases discoverability and readership?

To address this question and confirm a small number of hypotheses that derive from it, a randomly selected sample of eBooks newly published on Amazon in the first week of April, 2012, referred to as the Random Sample, was monitored for 15 weeks to measure the diffusion of information as captured by web searches and review counts along with sales inferred from sales rank. A second sample of the most popular eBooks (the Popular Sample) drawn at the same time as the Random Sample was also collected and tracked. The social media use of those authors was established and their social media participation compared with sales and book Web presence to determine to what extent social media participation predicts eBook sales and Web presence.

## **Summary of the Results with Conclusions**

This section provides a synopsis of the research findings detailed in Chapter 4 and conclusions that can be drawn from them.

### *Phase I Research - Discussion*

Phase I of the research consisted of drawing a random sample (referred to as the Random Sample) of newly published eBooks released by Amazon on their US site during the first week of April 2012, describing them and categorizing them for analysis. A non-random sample of popular titles (referred to as the Popular Sample) was also collected for comparison.

Initially, 483 titles were selected at random from a pool of 8,400, providing a margin of error of  $\pm 4.3\%$  at a confidence level of 95%. The non-random Popular Sample of eBooks included 190 titles.

The Random Sample included 58 foreign language titles, about 12% of the Random Sample. A thorough analysis was not undertaken, but a cursory review of the foreign language titles revealed a mix of content types including public domain works in the original language, public domain works translated from English to a foreign language, one or two bilingual dictionaries, some original fiction, one or two children's books, and one or two government reprints. These titles were not big sellers. Of the 47 that remained available by the end of the data collection period, nearly half recorded no sales, with the remainder selling only a few copies. However, eBooks are an economical way to produce titles for market that may not result in large sales. Amazon maintains separate sites in

foreign countries, so this is probably not representative of the worldwide output of eBook titles in languages other than English. For purposes of the remainder of the analyses, these titles were removed from the pool.

The re-publication of books in the public domain as new print on demand or eBooks has been noted in the literature (Bradley, Fulton and Helm, 2012; Bowker, 2012a), and by 2011, title output in this category had dwarfed all other categories, with millions of new issues of old titles according to Bowker, which tracks new titles through issuance of ISBN numbers. In the Random Sample, 62 titles were noted as falling into this “reprint jungle” category. While 14% is fairly substantial (and slightly higher than a random sample of self-published books from Bradley et al’s 2008 sample published in 2012), it is far less than would be expected had the 2011 trends continued to 2012 and into this Random Sample, indicating this trend may finally be slowing down or reversing. It may also be a result of Amazon removing these titles from the Kindle store (see Streitfeld, 2011). Several of the reprint titles disappeared before the research concluded. However, these titles will continue to cause bibliographic and consumer confusion in those cases where they are published as “new” titles with current publication dates.

The largest subcategory of books in the Random Sample were self-published works, at 73% of the Random Sample used for regression (62% of the total Random Sample not including foreign language titles). The latest data from Bowker (2012b) claims that self-published title releases were approximately equal to mainstream print title releases in 2012, but the Random Sample shows them to be higher than that by a much larger margin. Possible reasons for this include the observation that many eBook

titles and titles released only on Amazon and not in the general marketplace often lack ISBN numbers, which Bowker uses to track title production. There is also some confusion about how Bowker identifies self-published works. However if this number holds over the general population of eBooks, then self-published works comprise a greater segment of title production than is generally assumed.

The flip side of this is that self-published books from the Random Sample sold poorly compared to mainstream published books. In fact, 76 titles in the Random Sample failed to garner a single sale on Amazon, with the majority of those, 68, being self-published. On the other hand, some self-published books achieve top status. In the Popular Sample, nearly 40% of the titles were self-published including many titles initially or subsequently offered as free downloads on special promotions.

A majority of titles, 60%, in the Random Sample were absent from Barnes and Noble compared with only 14% of the titles from the Popular Sample. However, Amazon sometimes offers advantages to authors who agree to some period of exclusivity. This was not tested explicitly as a predictor since the details of the benefits are unknown, but it may be worthwhile to determine empirically whether releasing an eBook in multiple formats is a better strategy than remaining exclusive to one market channel.

Similarly, 80% of the titles in the Random Sample were eBook only, that is, lacking an available print version on the Amazon marketplace, compared to only 25% of the Popular Sample. Further, self-published books were far more frequently available as an eBook only (70% of the eBook only titles in the Random Sample were self-published). eBooks that also have traditional print availability (that is, mass printed and not print-on-

demand only) may benefit from certain marketing techniques used for print books, such as end-table displays in bookstores. That is, a person might discover the physical book and purchase the eBook. This research did not empirically investigate that aspect of discoverability, and this is an area for future research.

The original research proposal included plans for a review of titles based on subject, but this turned out to be difficult. Amazon uses its own system of key-words and hierarchies that will require a more refined approach to data collection than was possible using the available tools. Further, titles are not (were not) uniformly cataloged. For example, some instances of erotic fiction lacked that designation at the time the data was collected, but were later reclassified. Subject classification is part of the metadata shared by individuals and noted by reviewers, so the impact of proper classification of books on discoverability and sales should be a subject of future inquiry.

Although much has been made in the trade and popular press about enhanced eBooks and eBook applications, only a very small number in either the Random or Popular Sample were labeled as enhanced or optimized. This may reflect the Amazon e-paper versions of the Kindle eReader's lack of significant graphics or processing capabilities. For an empirical look at enhanced eBooks and eBook applications, a different pool of titles and title collection methods would need to be employed, and this also indicates that conclusions from this study based on a random sample from Amazon should be generalized only with caution. Apple, for example, has released many titles as applications, and provides a publishing platform, incompatible with Amazon Kindle, that authors can use to incorporate media more fully into eBooks and eBook applications.

Pricing from the Random Sample generally confirms what was reported in the trade and popular press. The median price for self-published eBooks in the Random Sample was \$2.99 with two modes at \$0.99 and \$2.99<sup>23</sup>. For mainstream published books, there was a single mode at \$9.99. Price continues to fluctuate for both traditionally published and self-published works, so this category bears watching and may turn out to be at least partially predictive of sales and diffusion.

Price manipulation is often recommended in the popular press as a technique for increasing discoverability on the retailers' sites. The technique supposedly works by lowering – or even giving away – a title for a promotional period. Sales rank improves as people take advantage of the promotion. The title rises in visibility because the recommendation engines and business analytics used by the retailers to increase sales favor rapid rise sellers. The popularity bounce continues for a period after the promotion ends thus ultimately improving discoverability, sales and profit. This hypothesis wasn't tested explicitly, but the data show that over one-third of the titles in the Random Sample used price manipulation, with the majority of those being self-published titles. This is an area that bears further empirical research.

Finally, large numbers of short fiction works released as eBooks, some as sparse as a single page ranging up to 20 or 30 pages in length, is a trend noted in the Random Sample, but not so much in the Popular Sample. Almost all self-published, a great number of the short titles were in the category of erotic romance or niche erotic fiction,

---

<sup>23</sup> These modes may result from Amazon's strategy of maximizing profit share paid to authors, which varies according to specific price points.

with others in categories as diverse as single one-page poems, recipes and even a quilting pattern.

Both the short fiction and erotic fiction categories were hard to classify, for different reasons. Page length is somewhat arbitrary on an eBook and Amazon doesn't report page length consistently. In some cases, word counts are given; in others, page length is estimated and in still others, no information is provided. File size cannot be used because even a small number of illustrations or color cover art increases file size disproportionately to the amount of text. Some reader reviews downgrade ratings for short titles not labeled as such.

The erotic fiction category was difficult to classify because of inconsistencies in subject key words. Although some short titles are clearly explicit, including some with explicit photography, Amazon does not provide an adults-only tag or filter, and some titles in the erotic romance category contain some explicit text. Cover art and description suggest, but do not confirm, definitive classification of many of these titles.

Although precise numbers were not attempted due to the difficulties in classification, it does appear that short fiction, erotic fiction, and short erotic fiction were not, in general, selling well from the Random Sample, although two of the best sellers in the Popular Sample were categorized as novel-length erotic fiction.

Short fiction and extended length non-fiction have been touted as an emerging literary form especially well-suited to the eBook. While both the Random and Popular Samples contained examples of both, the short form category did not appear to be among

the best-selling, and this finding contradicts some of the conventional wisdom. This category of short length eBooks deserves ongoing attention.

Finally, a breakdown of sales and sales rank by use of social media outreach shows clear differences between authors using at least some social media outreach and those who use none. 35% of the group that used no social media recorded no sales during the data collection period, vs. only 19% of those who did at least some social media outreach. Further, of those who recorded sales, use of social media increased the median sales number per period by 36% compared to those who used no social media outreach. This result pre-sages and is consistent with the results of the regressions described in the next section, and offers support for the social gatekeeping framework.

### Summary

Self-publishing dominates new title output on the Amazon Kindle format, in greater numbers based on the Random Sample than some market data can confirm, but comparatively more mainstream-published books found at least a limited market for downloads and sales. Although much has been written about the acceptance of self-publishing among the general reading public, the appearance of self-published titles on best-seller lists, and the defections to self-publishing of front- and mid-list authors formally published through the mainstream model, sales success is rare, at least as evidenced by sales figures from the Random Sample. On the other hand, few of the traditionally published titles in the Random Sample garnered more than limited numbers of sales over the short period studied and the data show that for at least the first several

weeks after publication, most titles are largely undiscovered or under-discovered.

However, the Phase II and Phase III study results, summarized in the pages that follow, provide evidence of some specific strategies authors might consider to improve their chances.

### *Phase II Research - Discussion*

Phase II of the research consisted of analysis of Author Web presence and social media participation as independent (predictor) variables, and book Web presence and sales as dependent variables. Multiple regression was the primary tool used to determine the degree to which the independent variables could predict variance in the dependent variables. The Social Gatekeeping framework developed in Chapter 1 suggests that author Web presence and participation in social media should associate positively with book Web presence and sales. The multiple regression analysis was used to show which specific author Web presence and social media activities might best predict sales and discoverability.

The dependent variables tested include search engine hit counts, Amazon sales rank and Amazon consumer review counts. For the search engine hit counts, two specific kinds of queries designed to provide results with high precision and high recall were tracked weekly for 15 weeks immediately following publication providing 16 counts for each title. One query was on the Amazon Stock Identification Number (ASIN), which is a unique 10 character alphanumeric string. A second query was on an exact phrase search of the form “*The-Book-Title* by *The-Book-Author*.” The ASIN query was conducted on Google and Bing. An additional ASIN query was conducted on a subset of Blog results on Google. The author-title query was conducted weekly on Bing. Google blocked automated retrieval of the author-title query, so it was conducted only once manually in the final week of data collection. Sales rank and review counts were retrieved weekly directly from Amazon.

Preparation of the dependent variables for analysis included weighted averaging of the final three counts to provide smoothing of noise, and transformation. Book sales are known from previous research to follow a power distribution. Data from the search engine hit counts and book reviews were subjected to curve fitting and also found to follow a power distribution. Therefore, a box-cox logarithmic transform algorithm was applied to the raw dependent variable data to derive data as nearly normally distributed as possible for the regressions.

Preparation of the independent variables largely included manual search and filter techniques. After elimination of titles from the Random Sample that did not come from living authors and titles that were otherwise unsuitable for an author analysis, such as magazines and government reprints, an author search on the remaining titles (N=325) of the Random Sample was conducted to determine if the author maintained a Web page, a blog, an amazon author page, a page on the social book network Goodreads, a Facebook page or a Twitter account. These were dichotomized, coded yes/no (1/0), and entered as independent variables in multiple regression analyses. The process was repeated to determine the independent variables for the Popular Sample.

A separate analysis was conducted to determine if there was contamination of the DV data with IV data, that is, whether the search engine hit counts included counts from the authors' web and blog sites. Contamination of the dependent variables (DV) with independent variable (IV) counts was found to be negligible.

Multiple regressions were computed on each of the DVs from both the Random Sample and the Popular Sample with the set of 6 IVs identified above to determine if

significant models developed. In all, 14 regressions were computed. One was discarded (Google search on ASIN on the blog pages subset) due to insufficient data and lack of normal distribution. Of the remaining regressions, all returned significant models with normal or near normal distributions of the residuals.

The following tables summarize the significant standardized Betas, which show consistency in predicting the dependent variables (refer to Chapter 4, Phase II, for an explanation of the regression statistics summarized here).

Table 26 - Summary of Betas, Random Sample

<b>Random Sample</b>						
DV:	Google ASIN	Google A/T	Bing ASIN	Bing AT	Amazon Sales	Amazon Reviews
R-Squared	<b>0.181</b>	<b>0.132</b>	<b>0.167</b>	<b>0.379</b>	<b>0.255</b>	<b>0.272</b>
Significant Betas:						
Amazon Author	0.185		0.114		0.135	
Goodreads	0.182	0.288	0.290	0.444	0.352	0.385
Facebook	0.154		0.150	0.161		0.225

Table 27 - Summary of Betas, Popular Sample

<b>Popular Sample</b>						
DV:	Google ASIN	Google A/T	Bing ASIN	Bing AT	Amazon Sales	Amazon Reviews
R-Square	<b>0.170</b>	<b>0.119</b>	<b>0.112</b>	<b>0.191</b>	<b>0.156</b>	<b>0.197</b>
Significant Betas:						
Amazon Author						0.164
Goodreads	0.100					
Web Page	0.162	0.167	0.168	0.161		
Facebook	0.243	0.185	0.206	0.302	0.296	0.295

In the Random Sample, the  $R^2$  values of the regressions ranged from a low of .132 to a high of .379, meaning that the use of social media model accounted for between 13% to nearly 40% of the variance observed in the dependent variables tested. Of the independent variables in the model, presence of a Goodreads page was the most predictive on each of the computed regressions, with a standardized Beta ranging from a

low of .182 to a high of .444, meaning that for every one standard deviation of increase in the independent variable, the presence of a Goodreads author page predicted rate of increase of between 18% and 44% of one standard deviation in the dependent variable. Of the other independent variables tested, a Facebook page had a significant Beta on four of the regressions and an Amazon author page had a significant beta on three of the regressions. Twitter, a Web page and a blog did not contribute significant betas to the regression model.

Significant models also developed for the Popular Sample regressions, but with a different set of significant predictors.  $R^2$  values were generally lower for the Popular Sample, ranging from .112 to .197 meaning that the use of social media accounted for between 11% and 20% of the variance observed in the dependent variables tested. Of the independent variables in the model, presence of a Facebook page was the most predictive, with significant Betas ranging from .185 to .302. Having a Web page, a twitter account or Blog did not contribute significant betas to the regression model.

### Discussion

The  $R^2$  values are generally moderate to low, but the use of social media by authors is not positioned as a complete model predicting all of the variance in book sales and Web presence. For example, the models do not take advertising, promotions, and other activities available to an author into account, nor do they take into account quality differences in the books themselves. Given that caveat, the results are significant and support the social gatekeeping framework and hypotheses derived from it.

The emergence of an author Goodreads page as the most consistent and significant predictor of reviews, sales and diffusion of information through the web for newly released titles provides strong support for the social gatekeeping framework. The Goodreads social network is comprised, according to their own website as of May 2013, of 17 million members who have added 550 million titles to their online bookshelves and created 23 million reviews. Goodreads members share reading as a common organizing interest, and members focus on reading recommendations, sharing reading interests with friends, and discovery of new reading materials based on social data and browsing of friends' bookshelves and lists.

Facebook, the next most predictive variable also relies on social sharing with a network of friends, although it is not limited to books. Both sites are places where persons of influence may find information about books, which is a prerequisite to sharing it with others and seeding the diffusion of information necessary for discoverability. Goodreads in particular, provides a high density of relevant book information that can be easily discovered through serendipitous browsing as well as focused searching, and of all the models, Goodreads' best predicts diffusion of information as evidenced by the high beta for the Bing search engine author-title model. Among new titles, Goodreads also predicts sales and Amazon reviews.

The relatively poor showing of Web and blog sites as predictors of sales and diffusion, at least when books are relatively unknown immediately following publication, may be an indication of the lack of suitability of general purpose search engines for serendipitous browsing. While books can easily be located if known by author and title,

there are few successful strategies for discovery of particular Web and blog sites based on keyword search given the large numbers of potential candidates, the limited number of returns provided by general purpose search engines, and their intended purpose of providing the most relevant results first, omitting less popular but equally relevant authors and titles. Similarly, persons already familiar with an author or title may follow an author on Twitter, but this doesn't seem to serve as a major catalyst for sales and the spread of information about new titles.

The  $R^2$  values computed for regressions on the Popular Sample are lower across the board than for the newly published Random Sample, and although the models are still significant, a different set of significant predictors emerge. There may be a few possible explanations; these would require further research to tease apart.

First, Goodreads almost drops off, appearing as a significant predictor on only one of the regressions. In contrast, Facebook is a consistent predictor across all models tested, along with Web pages, with the highest significant betas predicting Amazon sales and reviews. This may indicate that once a title reaches some critical mass, readers beyond Goodreads' motivated user base may seek out an author's Facebook and Web page for confirmation of information that ultimately leads to a purchase/read decision. That is, in contrast to Goodreads' influence in spreading information about a title through motivated and high engagement readers, Facebook and Web sites may play a more important role in driving sales once a title has come to the attention of a reader. Further research would be needed to confirm this interpretation, but it is consistent both with social gatekeeping and the diffusion of innovation frameworks that propose a multi-stage process of adoption

including diffusion of information followed by adoption based not only on information but on the opinions of friends and acquaintances in social networks.

Several caveats apply to the interpretation of this data. The first, and most important, is that the regression tool does not establish causality. That is, one can only say that the independent variable predicts variance in the dependent variable but not that it necessarily causes it. For example, it is possible that authors who have Facebook pages are more likely also to do some other thing, unidentified, that is the cause of the association. Different kinds of research would be necessary to support causality more robustly.

A second caveat is that the regressions only test author use of social media outreach, not social media generally. For example, an author having a Twitter account did not appear to be a strong predictor of eBook sales, but that doesn't mean that Twitter itself isn't a factor in increasing book Web presence or sales, only that author use of Twitter specifically didn't predict those variable.

Some caveats apply to the regression tool itself. One is that although many statistical tests are tolerant of datasets that are not normally distributed if they are reasonably close, the accuracy and precision of the calculated values, and the probability values ( $p$ ) decrease as the data departs from normality. Most of the data collected for this research, once transformed, was nearly, but not fully, normally distributed, although selected dependent variables did test normal after transformation. For regression, review of the residual values is informative. Appendix A contains a more complete discussion of

the individual datasets, normality, residual analysis and other statistical tests, for those with a background in regression.

A final caveat, or rather a set of related caveats, has to do with the use of regression in the social sciences generally. Regression is widely used across a wide range of disciplines to help determine the relative effectiveness of different kinds of strategies and interventions in areas as diverse as economics, education, finance, and social intervention. However, the accuracy and precision of both  $R^2$  and calculated Betas are sensitive to the proper development of a predictive model and selection predictor variables (Manzi, 2012). There is additional discussion of these factors in Chapter 2.

Omitted variable bias is perhaps the most damaging to interpretation of the Betas. Omitted variable bias occurs when one predictor variable correlates to a second predictor variable, but the second variable is not included in the regression. This has the tendency to give far more weight to the first predictor variable than is has in actual practice.

A second related issue arises when variables having a predictive effect are omitted from the regression even when not correlated to another variable used, that is, when regression is used to compute predictability for a partial model. If the regression is calculated with and without the omitted variable, it's both possible and likely that  $R^2$  will change, and the relative contribution of the individual predictors (the Betas) will also change. This is because the regression computes the variables as part of the whole model, not just the individual predictor variables. On the other hand, very complex models can be difficult to analyze using regression because as the number of predictor variables

increase, the number of cases necessary to provide sufficient power to the result increases as well.

Randomized field trials are the unquestionable gold standard for research involving the relative impact of factors on behaviors and outcomes, but these can be notoriously hard to do in the social sciences if there are many variables to control. Unlike aspirin tablets, for example, no two books are the same, nor are any two authors or publishers. To keep these properly managed in a randomized field trial could mean creating environments to control variables that would lack external validity.

For this research, the regressions are useful and supported both by other data collected and analyzed using different techniques (such as the clear difference in sales / no sales and sales rank among authors who did and did not use social media), and by popular and anecdotal accounts from industry analysts. So taken together, the methods in Phase I, II and III of this research suggest cautious confidence in the nature and direction of the results. But because many other factors undoubtedly affect the dependent variables in addition to the ones studied, the regression results should probably be tempered with realistic expectations of accuracy and precision.

In summary, the Phase II research empirically establishes a low to moderate association between author Web presence and social media on the one hand, and book Web presence and sales on the other, as evidenced by the significance of all computed regressions and specific coefficients of determination ( $R^2$ ). The results are consistent with the underlying theoretical components and affirm the hypotheses. The evidence presented here suggests that author Web presence and social media engagement are in fact

predictive of variance in the dependent variables. While these results do not establish causality, the results are consistent with the Social Gatekeeping framework. For authors (and publishers as well) the research at least tentatively suggests that outreach with media that puts the author in close contact with high engagement readers, especially including social reading sites, helps spread the word about books and triggers diffusion of information through social networks via the mechanism of social gatekeeping. However, once the basic information about a title comes to the attention of the reader, social engagement through general social network sites such as Facebook and maintenance of Web pages and other resources that provide descriptive and confirmatory information may be a factor in pushing the title from discoverability to sales.

### *Phase III Research - Discussion*

The purpose of the Phase III research was to conduct a review of a selected group of titles from the Random and Popular Samples in order to gain additional insight on how authors use social media and what such a review might suggest for future research. In all, about 35 titles were chosen based on observations made on initial review during data collection and also on results of the data collection such as the books ending data collection with the highest sales rank.

The results suggested, first of all that the independent variables were correctly chosen; that is, where social media and Web presence were used by an author, the categories selected as independent variables constituted the bulk of social media and Web presence use. This indicates that the omitted variable bias, at least with respect to social media engagement, was not a factor in the Phase II regressions. A handful of other social media venues were occasionally exploited by some authors, but for the most part, they were used infrequently by the majority of authors. Some bear watching as future trends. For example, video trailers are becoming more common even among low budget self-published authors.

There is considerably more variation in how and to what extent social media are used by authors, and this may account for at least some of the variance that was not accounted for by the regression models, as evidenced by the moderate to low  $R^2$  values. However, no clear picture emerged of how social media use might be better categorized and differentiated. Two measures were recorded during data collection, the actual number of Facebook friends or likes, and for Twitter users, the number of tweets and followers.

An ad hoc regression was run showing a positive model for number of tweets, but the beta was not large and was overshadowed by the other factors. No significant model emerged for the number of Facebook friends and likes. So this approach awaits refinement for future study.

One trend was noted in comparing Web sites as initially reviewed and again as they existed a year later. The use of widgets that encourage visitors to share on social sites appears to be increasing. These not only encourage sharing but also can provide web site administrators with quantitative data about how and where information is being shared. These touch points represent at least one manifestation of the serendipitous tie, that is, the sharing of information via a social contact that is not in one's own social network, even weakly. The serendipitous tie falls out of natural considerations of the social gatekeeping framework and harmonizes library and information science-based research on the importance of serendipity to browse behavior on the one hand, and the communication and sociology research on the importance of social ties and the diffusion of information, on the other. The Share widgets, which are becoming more sophisticated at tracking and monitoring sharing behaviors, may be a source of data for future research on this topic, along with other emerging techniques such as the tracking of the sharing of book specific URLs (for example, when the book's Amazon page URL is the actual message shared in a social context).

The final note of interest uncovered in Phase III was the number of publishers making specific references to social media outreach by authors on the publisher's submission guidelines pages. The data show that even among the titles published by the

top traditional model publishers, authors are engaging with social media, and more so among the top selling authors than from a random sample of authors.

### **Significance of the Research**

Reading and the cultural production of literary works are at risk without an effective way of connecting readers to books. As the nature of the book itself changes as text migrates to digital form and authors increasingly seek non-traditional paths to reader discovery and reception, traditional gatekeepers and the mass media represent only a small number of channels through which readers come to discover books. This research informs key stakeholders in the business and art of book culture of the changing nature of the reader-author connection, the emerging role of the author in connecting books with readers and the role of social networks in facilitating discovery and retrieval. It also progresses gatekeeping theory to accommodate new social network conceptualizations and lays a necessary foundation for ongoing research into the study of the emerging digital book market in coming years.

The impact of the social gatekeeping framework, and the introduction of the serendipitous tie conceptually introduced for this research, extends even beyond considerations of the book trade and its environs as outlined here. While social network theory and descriptions of the mechanics of information diffusion through social networks are extensive and well developed, social gatekeeping introduces a new way of thinking about the individual processes catalyzing the diffusion of information not only within social networks but between and through them, based on an extensive tradition of gatekeeping theory that has not previously been applied to individuals in social networks. The role of the individual as social network filter and the identification of the message as a base unit of analysis opens the door to new ways of thinking about how individuals and

social networks create, in effect, a mass media construct that can be understood and empirically analyzed through the lens of social gatekeeping.

### **Implications for Future Research and Recommendations**

While this research produced positive results supporting the hypotheses put forward, the results are specific to the circumstances of the data collected and can only be generalized with caution. In particular, the Random Sample came from a single source, Amazon, which may not necessarily be representative of the total population of books from all publishers. The problem is that there are no single sources of bibliographic control and no source from which a true random sample of the total pool of published eBooks can be drawn. This may be an intractable problem, but the results of this research should be extended to titles drawn from other pools and compared in order to generalize the results with greater confidence.

There was good but not perfect correlation among the dependent variables, implying that while they appear to be generally consistent in serving as a proxy for Web diffusion, reviews and sales, they measure different portions of the total number of theoretically relevant messages. These DVs should be further analyzed and refined to determine if confidence in their suitability as variables is fully warranted, or whether some other proxy can be found that is more suitable.

This research did not use experimental methods such as randomized field trials. Rather, these results provide the initial empirical foundations of support for extending Gatekeeping Theory to include Social Gatekeeping as an important construct that should

be further explored and validated using more costly and complex experimental methods. As such, it establishes a research agenda that can be progressed using increasingly sophisticated methods and tools.

One of the objectives of this research was to capture a snapshot of titles that could be examined and compared with other snapshots taken over time. There is considerably more data that could be mined from these data alone that was beyond the scope of this research.

High on the list of research questions that might be proposed for future research include reviewing libraries as a source of discovery and information diffusion, that is, their potential role as an independent variable predicting Web diffusion and sales, and also their role as a dependent variable as a measure of Web diffusion. Libraries and the role libraries play in both bibliographic control of digital titles and the degree and manner to which they make digital titles available is an evolving issue that could be informed by empirical research.

As this study is concluding, Amazon announced their intent to purchase Goodreads and had previously acquired an interest in Shelfari and Librarything, also book-centered social networks. What the outcome of this will be is unknown at the time of this writing, but it is clear that there is keen interest and awareness by commercial interests in social networks in general. For the book trade in particular, the growing importance of social networks as a catalyst for reading and the ongoing evolution of the book in both print and digital formats should prompt continued scholarly examination of

the role social networks and the Social Gatekeeping framework play in connecting readers to books.

More generally, the social gatekeeping framework and the role of the serendipitous tie in propagation of information through networks should be explored in other contexts to determine how, to what extent, and in what ways they are generalizable.

## **APPENDIX A - INDIVIDUAL REGRESSION RESULTS – COMPLETE**

The following entries provide the complete description of the regression data from Chapter 4 including description of the Random and Popular Samples and a discussion of the analysis of the residuals. All dependent variables were transformed prior to analysis using a Box-Cox logarithmic transformation computed using the SPSS built-in function. An understanding of multivariate regression and the individual tests and descriptions of regression analysis is presumed.

*DV = Google search engine counts on ASIN, Random Sample*

The distribution of the transformed dependent variable was found to be near normal with a standardized mean of near 0 and a standard deviation of 1.00, with skew of .517 and kurtosis of -.483. The Kolmogorof-Smirnov test was significant at  $p=.000$ , however visual examination of the histogram and Q-Q plot are free of obvious deviation from the normal distribution and free of obvious outliers.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .181$ ,  $F(6,318) = 11.711$ ,  $p < .001$ . Three of the independent variables had significant Betas as follows:

## Google ASIN Betas, Random Sample

Predictor	Standardized Beta	Significance
Amazon Author Page	.185	.001
Goodreads Author Page	.182	.002
Facebook Page	.154	.015

Standardized Betas for Web page, Blog and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.874, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) appear normal with no obvious signs of heteroskedasticity in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is near normal with a mean of near 0 and a standard deviation of .991, skew of .333 and a kurtosis of -.347. The Kolmogorof-Smirnov test was significant at  $p=.000$ , however visual examination of the histogram and Q-Q plot of the residuals are free of obvious deviation from a normal distribution and free of obvious outliers.

The regression was run a second time using the sub-population of self-published titles. The model was significant with  $R^2 = .105$ ,  $F(6,230) = 4.474$ ,  $p<.001$ . Of the predictor values, only the Amazon author page had a significant standardized Beta at .145,  $p=.029$ .

By comparison, when the sub-population was limited to mainstream published titles, the model was significant with  $R^2 = .255$ ,  $F(6,81) = 4.621$ ,  $p < .001$ . Of the predictor values, only the Amazon author page had a significant standardized Beta at .249,  $p = .024$ .

*DV = Google search engine counts on ASIN, Popular Sample*

The process was repeated for the non-random Popular Sample.

After removal of an obvious outlier, the distribution of the transformed dependent variable was found to be normal with a standardized mean of .03 and a standard deviation of .905, negative skew of .075 and positive kurtosis of .920. The Kolmogorof-Smirnov test was not significant at  $p=.200$ , and visual examination of the histogram and Q-Q plot are free of obvious deviation from the normal distribution and free of obvious outliers.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .170$ ,  $F(6,171) = 5.847$ ,  $p < .001$ . Three of the independent variables had significant Betas as follows:

Google ASIN Betas, Popular Sample

Predictor	Standardized Beta	Significance
Goodreads Author Page	.100	.047
Web Page	.162	.037
Facebook Page	.243	.002

Amazon author page, blog and Twitter account were not significant predictors.

There was no evidence of collinearity among any of the variables. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) appear normal with only slight evidence of heteroskedasticity in a plot of standardized predicted value vs. studentized residual value. The distribution of the residual is normal with a standardized mean of near 0 and a standard deviation of .983, negative skew of .104 and a positive kurtosis of .542. The Kolmogorof-Smirnov test was not significant at  $p=.200$ , and visual examination of the histogram and Q-Q plot of the residuals are free of obvious deviation from a normal distribution and free of obvious outliers.

*DV = Google search engine counts, ASIN in Blog Pages, Random Sample*

For the Random Sample, the frequency table of calculated weighted moving averages of the dependent variable showed that nearly 95% of cases in the Random Sample had 1 or fewer average counts, meaning the search found almost no penetration of the ASIN in the blog space as reported by Google. Although regression reported a significant model, with very low  $R^2$ , neither the transformed dependent variable nor the residuals were deemed normal or sufficient enough to consider the results suitable for regression.

*DV = Google search engine counts, ASIN in Blog Pages, Popular Sample*

After removal of an outlier, the distribution of the transformed dependent variable was found to be near normal with a standardized mean of .02 and a standard deviation of .997, with skew of .153 and kurtosis of 1.165. The Kolmogorof-Smirnov test was significant at  $p=.000$ , however visual examination of the histogram and Q-Q plot are free of major deviation from the normal distribution and free of obvious outliers, with the exception of more cases than would be expected at the extreme tails.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .283$ ,  $F(6,161) = 2.479$ ,  $p = .025$ .

Of the predictor variables, only Facebook had a significant beta of .204,  $p=.015$

Standardized Betas for the other predictors were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.723, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) appear normal, however with evident heteroskedasticity, in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is near normal with a mean of near 0 and a standard deviation of .983, skew of -.120 and a kurtosis of 1.485. The Kolmogorof-Smirnov test was significant at  $p=.001$ , however visual examination of the histogram and Q-Q plot of the

residuals are free of obvious deviation from a normal distribution and free of obvious outliers.

*DV = Google search engine counts on Author - Title, Random Sample*

The distribution of the transformed dependent variable was found to deviate from normal distribution with a standardized mean of near 0 and a standard deviation of 1.00, with skew of -.480 and kurtosis of -1.171. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot shows a large pool of cases at the very low end of the scale.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .132$ ,  $F(6,318) = 8.060$ ,  $p < .001$ .

Only the Goodreads predictor had a significant standardized beta of .288,  $p < .001$

Standardized Betas for the other predictors were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.845, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) appear skewed with some evidence of heteroskedasticity in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is deviates from normal with a mean of near 0 and a standard deviation of .991, skew of -.609 and a kurtosis of -.774. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot of the residuals deviate moderately from a normal distribution, with a fat left tail and visible negative skew; the plots are free of obvious outliers.

*DV = Google search engine counts on Author - Title, Popular Sample*

After removal of an outlier, the distribution of the transformed dependent variable was found to be near normal with a standardized mean of .03 and a standard deviation of .934, with skew of .494 and kurtosis of .011. The Kolmogorof-Smirnov test was significant at  $p=.004$ , however visual examination of the histogram and Q-Q plot are free of obvious deviation from the normal distribution and free of obvious outliers.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Google search engine hit counts, with  $R^2 = .119$ ,  $F(6,161) = 3.839$ ,  $p = .001$ . Two of the independent variables had significant Betas as follows:

Google A/T Betas, Popular Sample

Predictor	Standardized Beta	Significance
Web page	.167	.036
Facebook	.185	.024

Standardized Betas for the remaining predictors were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.671, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) appear normal but with evidence of heteroskedasticity in a plot of

standardized predicted values vs. studentized residuals. The distribution of the standardized residual is normal with a mean of near 0 and a standard deviation of .983, skew of .287 and a kurtosis of .103. The Kolmogorof-Smirnov test was not significant at  $p=.200$ , and visual examination of the histogram and Q-Q plot of the residuals are free of obvious deviation from a normal distribution and free of obvious outliers.

*DV = Bing search engine counts on ASIN, Random Sample*

The distribution of the transformed dependent variable was found to be not normally distributed by virtue of over 54% of the cases having a value of 0 indicating no search engine counts. The standardized mean was near 0 and the standard deviation was 1.00, with skew of .354 and kurtosis of -1.698. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot confirmed that the transformed dependent variable was not normally distributed.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Bing search engine hit counts, with  $R^2 = .167$ ,  $F(6,318) = 10.625$ ,  $p < .001$ . Three of the independent variables had significant Betas as follows:

Bing ASIN Betas, Random Sample

Predictor	Standardized Beta	Significance
Amazon Author Page	.114	.038
Goodreads Author Page	.290	.000
Facebook Page	.150	.019

Standardized Betas for Web page, Blog and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.914, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) do not appear normal and there are signs of heteroskedasticity in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is not normal in spite of a mean of near 0 and a standard deviation of .991, skew of .281 and a kurtosis of -1.202. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot of the residuals indicate deviation from a normal distribution; the plots are free of obvious outliers.

*DV = Bing search engine counts on ASIN, Popular Sample*

The distribution of the transformed dependent variable was found to be not normally distributed. The standardized mean was near 0 and the standard deviation was 1.00, with skew of .014 and kurtosis of -.544. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot confirmed that the transformed dependent variable was not normally distributed, with two clear modes appearing on the plots.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Bing search engine hit counts, with  $R^2 = .112$ ,  $F(6,172) = 3.633$ ,  $p = .002$ . Two of the independent variables had significant Betas as follows:

Bing ASIN Betas, Popular Sample

Predictor	Standardized Beta	Significance
Web Page	.168	.035
Facebook	.206	.012

Standardized Betas for Amazon page, Goodreads page, Blog and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.632, indicates that the residuals are not serially correlated. Visual

examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) do not appear normal and there are signs of heteroskedasticity in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is not normal in spite of a mean of near 0 and a standard deviation of .983, skew of -.262 and a kurtosis of -.385. The Kolmogorof-Smirnov test was significant at  $p=.001$ , and visual examination of the histogram and Q-Q plot of the residuals indicate deviation from a normal distribution; the plots are free of obvious outliers.

*DV = Bing search engine counts on Author – Title, Random Sample*

The distribution of the transformed dependent variable was found to be not normally distributed by virtue of 52% of the cases having a value of 0 indicating no search engine counts. The standardized mean was near 0 and the standard deviation was 1.00, with skew of .967 and kurtosis of -.354. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot confirmed that the transformed dependent variable was not normally distributed.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Bing search engine hit counts, with  $R^2 = .379$ ,  $F(6,318) = 32.333$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Bing A/T Betas, Random Sample

Predictor	Standardized Beta	Significance
Goodreads Author Page	.444	.000
Facebook Page	.161	.003

Standardized Betas for Amazon Web page, Blog and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.683, indicates that the residuals are not serially correlated. Visual

examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) appear close to normal but with some evidence of heteroskedasticity in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is near normal with a mean of near 0 and a standard deviation of .991, skew of .235 and a kurtosis of .205. The Kolmogorof-Smirnov test was significant at  $p=.000$ , however visual examination of the histogram and Q-Q plot of the residuals are free of obvious deviation from a normal distribution, with the exception of a sharp central peak, and free of obvious outliers.

*DV = Bing search engine counts on Author – Title, Popular Sample*

The distribution of the transformed dependent variable was found to be not normally distributed. The standardized mean was near 0 and the standard deviation was 1.00, with skew of .40 and kurtosis of -.917. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot confirmed that the transformed dependent variable was not normally distributed.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Bing search engine hit counts, with  $R^2 = .191$ ,  $F(6,172) = 6.757$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Bing A/T, Popular Sample

Predictor	Standardized Beta	Significance
Web page	.161	.034
Facebook Page	.302	.000

Standardized Betas for Amazon Web page, Goodreads page, Blog and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.576, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted

cumulative probability) appear close to normal but with some evidence of heteroskedasticity in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is near normal with a mean of near 0 and a standard deviation of .983, skew of -.286 and a kurtosis of -.486 The Kolmogorof-Smirnov test was significant at  $p=.000$ ; visual examination of the histogram and Q-Q plot of the residuals deviate somewhat from a normal distribution, with modal peaks at  $\pm 1$  SD; there are no obvious outliers.

*DV = Amazon Sales, Random Sample*

The distribution of the transformed dependent variable was found to be not normal with nearly 25% of the Random Sample at zero sales. Results show a standardized mean of near 0 and a standard deviation of 1.00, with skew of .129 and kurtosis of -.951. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot differ from the normal distribution.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Amazon sales, with  $R^2 = .255$ ,  $F(6,315) = 17.959$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Amazon Sales Betas, Random Sample

Predictor	Standardized Beta	Significance
Amazon Author Page	.135	.010
Goodreads Author Page	.352	.000

Standardized Betas for Web page, Blog, Facebook and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.759, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted

cumulative probability) appear nearly normal but with signs of heteroskedasticity in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is near normal with a mean of near 0 and a standard deviation of .991, skew of -.063 and a kurtosis of -.621. The Kolmogorof-Smirnov test was significant at  $p=.000$ , however visual examination of the histogram and Q-Q plot of the residuals appear to have nearly normal distribution and are free of obvious outliers.

*DV = Amazon Sales, Popular Sample*

The distribution of the transformed dependent variable was found to be not normal with a positively skewed appearance. Results show a standardized mean of near 0 and a standard deviation of 1.00, with skew of .564 and kurtosis of -.446. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot differ from a normal distribution.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Amazon sales, with  $R^2 = .156$ ,  $F(6,171) = 5.273$ ,  $p < .001$ . One of the independent variables had significant Betas as follows:

Amazon Sales Betas, Popular Sample

Predictor	Standardized Beta	Significance
Facebook	.296	.000

Standardized Betas for Goodreads page, Amazon page, Web page, Blog, and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.776, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) appear normal but with signs of heteroskedasticity in a plot of

standardized predicted values vs. studentized residuals. The distribution of the standardized residual is near normal with a mean of near 0 and a standard deviation of .983, skew of -.435 and a kurtosis of -.019. The Kolmogorof-Smirnov test was not significant at  $p=.052$ , and visual examination of the histogram and Q-Q plot of the residuals appear to have normal distribution and are free of obvious outliers.

*DV = Amazon Review Count, Random Sample*

The distribution of the transformed dependent variable was found to be not normal with over 65% of the Random Sample lacking a single review. Results show a standardized mean of near 0 and a standard deviation of 1.00, with skew of .789 and kurtosis of -1.231. The Kolmogorof-Smirnov test was significant at  $p=.000$ , and visual examination of the histogram and Q-Q plot differ markedly from a normal distribution.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Amazon review count, with  $R^2 = .272$ ,  $F(6,315) = 19.639$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Amazon Reviews Betas, Random Sample

Predictor	Standardized Beta	Significance
Goodreads page	.385	.000
Facebook page	.225	.000

The Standardized Beta for Amazon page was nearly significant at .100,  $p=.051$ . Standardized Betas for Web page, Blog, and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 2.029, indicates that the residuals are not serially correlated. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted

cumulative probability) appear nearly normal with a strong peak at the 0 SD mark but with signs of heteroskedasticity in a plot of standardized predicted values vs. studentized residuals. The distribution of the standardized residual is near normal with a mean of near 0 and a standard deviation of .991, skew of .477 and a kurtosis of -.422. The Kolmogorof-Smirnov test was significant at  $p=.000$ , however visual examination of the histogram and Q-Q plot of the residuals appear to have nearly normal distribution, with the exception of a strong central spike, and are free of obvious outliers.

*DV = Amazon Review Count, Popular Sample*

The distribution of the transformed dependent variable is normal. Results show a standardized mean of near 0 and a standard deviation of 1.00, with skew of .338 and kurtosis of -.433. The Kolmogorof-Smirnov test was not significant at  $p=.200$ , and visual examination of the histogram and Q-Q plot appear normal.

Using the Enter method, a significant model emerged indicating that author web presence as measured by the independent variables significantly predicts Amazon review count, with  $R^2 = .197$ ,  $F(6,171) = 7.012$ ,  $p < .001$ . Two of the independent variables had significant Betas as follows:

Amazon Reviews Betas, Popular Sample

Predictor	Standardized Beta	Significance
Facebook page	.295	.000
Amazon page	.164	.029

Standardized Betas for Goodreads page, Web page, Blog, and Twitter account were not significant.

There was no evidence of collinearity among any of the variables. The Durbin-Watson result, at 1.441, indicates borderline serial correlation. Visual examination of the residual plots (histogram, P-P plot of observed vs. predicted cumulative probability) appear nearly normal but with signs of heteroskedasticity in a plot of standardized

predicted values vs. studentized residuals. The distribution of the standardized residual is near normal with a mean of near 0 and a standard deviation of .983, skew of .218 and a kurtosis of -.137. The Kolmogorof-Smirnov test was not significant at  $p=.200$ , and visual examination of the histogram and Q-Q plot of the residuals appear to have normal distribution and are free of obvious outliers.

## APPENDIX B – RANDOM SAMPLE TITLES

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B004U7G0MG	celebraTORI	Tori Spelling	Gallery Books	n	5	y
B0053SGVO8	Wohnmobil Reisebuch ThermalbÄnder um Bad Buchau (German Edition)	Karin Ulrike Gerkhardt	Terranautic World Limited			n
B00558VHPY	Create A Stunning Freshwater Aquarium (Bringing Joy to your life through the magic of companionship)	Ron Mahon	Inmrc - Publish	y	2	n
B0055OIE2C	Black Heart (Curse Workers)	Holly Black	Margaret K. McElderry Books	n	6	y
B005FFULVS	Caring Is Creepy	David Zimmerman	Soho Press	n	3	y
B005FLODDE	The Woman Who Wasn't There	Robin Gaby Fisher	Touchstone	n	5	y
B005GG0MV0	Gypped	Carol Higgins Clark	Scribner	n	3	y
B005GSYXGI	So You Created a Wormhole: The Time Traveler's Guide to Time Travel	Phil Hornshaw	Berkley	n	5	y

---

<sup>24</sup> Number of Amazon page, Goodreads page, Facebook page, Web page, Blog, Twitter account

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B005GSYXXQ	Rurally Screwed: My Life Off the Grid with the Cowboy I Love	Jessie Knadler	Berkley	n	5	y
B005GSZHYA	Just Down the Road	Jodi Thomas	Berkley	n	5	y
B005GSZZ2Y	Sweet Addiction	Maya Banks	Berkley	n	5	y
B005RZB5IO	The Mammoth Book of Steampunk (Mammoth Books)	Sean Wallace	Robinson	n	4	y
B005SJLUHU	La Duende Vampira: Amor, Guerra y Tristeza Libro 1 (vampiros - series de vampiros - romance paranormal - mitologia - espa�ol) (Spanish Edition)	Vianka Van Bokkem				n
B006JUV7Z2	Andrade (Spanish Edition)	Esteban Navarro	Esteban Navarro			n
B006QBDKQS	The Impossible State: North Korea, Past and Future	Victor Cha	Ecco	n	2	y
B006YAD0CA	Sydney Harbor Hospital: Luca's Bad Girl	Amy Andrews	Harlequin Medical Romances	n	3	y
B006YAD1A6	Tamed by a Texan (Harlequin American Romance)	Tanya Michaels	Harlequin American Romance	n	4	y
B006YAD87M	The Road to Three Creeks (Harlequin	Caron Todd	Harlequin Special Releases	n	1	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
	Heartwarming)					
B006YADBW4	The Pretend Proposal (Harlequin Romance)	Jackie Braun	Harlequin Romance	n	5	y
B006YADDAE	Wicked	Crystal Jordan	Spice Briefs	n	5	y
B006YADDJO	The Ex Who Hired Her (Harlequin Presents Extra)	Kate Hardy	Harlequin Presents	n	5	y
B006YADDRC	The Widow's Protector (Love Inspired Suspense)	Stephanie Newton	Love Inspired Suspense	n	5	y
B00702M4YG	Fall of Interpretation, The	James K. A. Smith	Baker Academic	n	5	y
B00702M59U	Heaven Is Now	Andrew Farley	Baker Books	n	4	y
B0072O00P2	Banner of the Damned	Sherwood Smith	Daw	n	5	y
B00739NHVA	JoJo und ich: Die Geschichte einer tiefen Freundschaft (German Edition)	Dean Bernal	Integral			n
B0073UN82W	Discovery, The	Dan Walsh	Revell	n	6	y
B0073XV6O6	The Sign: The Shroud of Turin and the Secret of the Resurrection	Thomas de Wesselow	DUTTON ADULT	n	1	y
B0076M5LD6	The Daughters of Gentlemen: A Frances Doughty Mystery (Frances Doughty Mystery 2)	Linda Stratmann	The History Press	n	5	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B0078XH4IS	The 10 Best Decisions a Graduate Can Make	Bill Farrel	Harvest House Publishers	n	3	y
B00796LM1O	The Diviner: The inspiring true story of a man with uncanny insight and the ability to heal	Joe Cassidy	Penguin	n	4	y
B007B2XYU8	Oral Pathology E-Book: A Comprehensive Atlas and Text with EXPERT CONSULT - Online and Print	Sook-Bin Woo	A Saunders Title	n	1	y
B007C73A7E	RedBone	T. Styles	Urban Books	y	6	y
B007CJJBXY	Fire Child, Water Child: How Understanding the Five Types of ADHD Can Help You Improve Your Child's Self-Esteem and Attention	Stephen Cowan	New Harbinger Publications	n	3	y
B007CJJC0G	Healing the Angry Brain: How Understanding the Way Your Brain Works Can Help You Control Anger and Aggression	Ronald Potter-Efron	New Harbinger Publications	n	1	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007EF7VUG	The Law of Agreement: Discover the True Power of Intention	Tony Burroughs	Weiser Books	n	5	y
B007EFHQHE	The Mother Road	Jennifer ALee	Abingdon Press	n	6	y
B007FFKV9S	Cast of Characters	0	The Fiction Studio	y	4	y
B007FUMPSI	The Grey Among The Green	John Fuller	Vintage Digital	n		n
B007GAYPCG	Ready, Scrap, Shoot (A Kiki Lowenstein Scrap-N-Craft Mystery)	Joanna Campbell Slan	MIDNIGHT INK	n	6	y
B007HLYJES	The God Who Sees You: Look to Him When You Feel Discouraged, Forgotten, or Invisible	Tammy Maltby	David C. Cook	n	5	y
B007IVBN3W	A Topps League Story: Book One: Jinxed! (Topps Town)	Kurtis Scaletta	Amulet Books	n	5	y
B007JC1ZEW	Experiencing Spiritual Breakthroughs	Bruce Wilkinson	Multnomah Books	n	5	y
B007JK59YQ	The Marriage Assignment	Lynn Patrick	Harlequin Special Releases	n	2	y
B007JLRZHY	The Swordsman Chronicles (Swordsmen Chronicles)	Edmund Sim		y	0	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007K1EE8G	The Black Hole of Empire: History of a Global Practice of Power	Partha Chatterjee	Princeton University Press	n	3	y
B007KZXYVK	Alexander The Great	Graham Phillips	Virgin Digital	n	3	y
B007L5CIXY	Dora va en busca del las estaciones (Spanish Edition)	Samantha Berger	Nickelodeon Publishing			n
B007MAU9XE	My Imaginary Jesus	Matt Mikalatos	Tyndale House Publishers, Inc.	n	4	y
B007MEUTS0	The Ethics of Business: A Zondervan Digital Short	Zondervan	Zondervan	n	1	y
B007MEUVYM	Winning over Your Emotions (Sandy's Tea Society)	H. Norman Wright	Harvest House Publishers	n	2	y
B007N6ZJT6	Carlos: Portrait of a Terrorist: In Pursuit of the Jackal, 1975-2011	Colin Smith	Penguin	n	4	y
B007NACVAW	The Uncanny X-Men: An Origin Story	Rich Thomas	Marvel Press ebook	n	1	y
B007NZQGO4	The Red House Mystery (Vintage Classics)	A. A. Milne	Vintage Digital	n		n
B007O04CZI	Preparing for the AP Biology Examination: Fast Track To A 5, 1st Edition	Robert Doltar	Delmar Learning	y	1	n

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007O04GC2	Fitness!, 5th Edition (Cengage Learning Activity)	Karen S. Mazzeo	Brooks/Cole	n	1	n
B007O36VMW	Top 10 Montreal & Quebec City (EYEWITNESS TOP 10 TRAVEL GUIDE)	Gregory Gallagher	DK Publishing	n	3	y
B007OWPOF8	Empty Promises Participant's Guide	Pete Wilson	Thomas Nelson	n	4	y
B007P05AJE	Marc Morrone's Ask the Bird Keeper (Marc Morrone Pets Series)	Amy Fernandez	BowTie Press	n	3	y
B007P2823S	Micro Multinationals: A guide to international finance for small businesses	Emily Coltman	Harriman House	n	1	y
B007Q1H1IK	Strange Powers, Stranger Places	Michael Angel	Banty Hen Publishing	y	3	y
B007Q27VW0	The Incomers	Moira McPartlin	Fledgling Press	n	4	y
B007QGE1I8	Farm Animals: Picture Book (Educational Children's Books Collection) (Planet Collection)	Planet Collection	Planet Collection	y	0	y
B007QGE200	The House of Cards Trilogy	Barbara Metzger	Untreed Reads Publishing	n	3	y
B007QGNJ4K	Tickled Pink	Tessa Wanton	Naughty Nights Press	n	4	y
B007QGNJV8	The Studio Floor	Rose Rancourt		y	1	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QGNMTC	The Evil Queen & The Two Princesses (JAFF Just Another Friggin' Fairytale)	Grimy Brothers	Red Kitty's Publishing	y	0	y
B007QGQICU	99 Tipps zum iPad - Die besten Profi-Tipps für iPad-Nutzer (German Edition)	Wilfred Lindo				n
B007QGX0U8	Zog Land	Harry Roldan		y	0	n
B007QH286Y	The Sweetest Touch (Brothers of Worthington)	Marie Higgins	Canyonland Press	y	3	y
B007QHJJBQ	Brooklyn Bites: A Pickle & Carrot Cake	Scott Stabile	McCarren Park Publishing	y	4	y
B007QI2FAC	Blood Brothers Book Two: Warrior's Journey	Sadie and Sophie Cuffe	Desert Breeze Publishing	y	2	y
B007QI4U5K	Cul-De-Sac: A Suburban Erotica In Around	Elliot Silvestri	Green Bush Publishing	y	2	y
B007QI4VRC	The Quiet Girl	JJ Argus		y	4	y
B007QIB4BS	Best Russian Short Stories: selected novella from 19th Century (Illustrated + free audiobook version)	THOMAS SELTZER	Gentlehand Press	n		n
B007QIB6M0	The Slave Game	JJ Argus		y	4	y
B007QIBIKA	Deliverables	Tom Spears	Tom Spears	y	5	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QIMI7M	One Night of Lust	Zoe Waters		y	0	y
B007QIYGXG	EASY Thai FOOD For Foreigner	Chef Closer	Chef Closer	y	0	y
B007QK986E	Captivated	Leen Elle		y	2	y
B007QKINHE	Tommy and the Easter Egg Hunt (Tommy the Lifeboat Bear)	Beverly Lucas-Brown		y	1	y
B007QL5WRM	Tearing feelings (Compulsion of desire)	Kelvin Waiden		y	1	y
B007QM4056	Orpheus: Myths of the World	Padraic Colum		n		n
B007QM409M	The Last Woman on Earth	Han Drimmer		y	3	y
B007QM9NFS	Hot Water	Donald Campbell		y	1	y
B007QMCRIS	Five Tales from the Workshop at the End of the World with Kris 'n' Dean	Alistair Ainscott	Rapid-Dynamix Publishing	y	2	y
B007QMEAGA	Jewel of Classical Horror Novels Collection; Dracula & Dracula's Guest (Annotated) by Bram Stoker, Frankenstein (Annotated) by Mary Shelly	Bram Stoker		n		n
B007QMECLI	Erotic Story - The Bachelor Party	Sandra P		y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QMHQPC	The Point of View (Annotated with Biography of the Author)	Henry James		n		n
B007QMN7JQ	Rasselas Prince of Abyssinia:[Illustrated]	Samuel Johnson		n		n
B007QMP2IK	The Drums of Fu Manchu	Sax Rohmer		n		n
B007QMPHKI	EE Words - Colour (Phonic Words Colour)	Billy Bumble	JayCurt LLP	y	4	y
B007QN009U	From Ho to Housewife (My testimony) Volume 1	Shakea Sanders				n
B007QN7SZ4	Clown, the Circus Dog : Clown's Puppy Days(Full Illustrated)	A. Vimar	Orange jasmine	n		n
B007QNPVZ8	Bootstrap Marketing: 101 Top Tips for Marketing Your Business on a Budget	Mike Morrison	Bootstrap Publishing	y	2	y
B007QNUGQC	Building a Better Mouse Trap: Increasing Counter Terrorism Capabilities through Consolidation	Christopher O. Vicino		y	0	y
B007QNULDU	Amish Baby Quilt Pattern	Anniken Davenport	Valhalla Press	y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QNVL8Y	Poetically Correct: Banned by The Free Press	Ciera S. Louise	Trafford	y	2	y
B007QNVUZI	MARVEL Mansion Gang	Connie Ellen	Trafford	y	1	y
B007QO45OU	Cyber (Spanish Edition)	Julio Garc�a Castillo				n
B007QO49FK	The Mission of Alexis Dering (The Novella Series)	Laura Joyce Moriarty		y	2	y
B007QOD80C	Someday I'll Find You	Kate Sweeney	Intaglio Publications	n	3	y
B007QOFURG	Love's Shelter	LeichelleK		y	5	y
B007QOH4L6	The Evelyn Project	Kfir Luzzatto	Pine Ten, LLC	y	4	y
B007QOH73Q	Bullies Don't Have Game (The Little Rednecks and a Town Full of Bullies)	T Denise Robinson	Ragz Books	y	4	y
B007QOI3FM	Easy selections from Plato	Andromeda Publishing	Andromeda Publishing	n		n
B007QOJLU8	DUIVELSDREK - Twaalf korte verhalen - (Dutch Edition)	Hans van Cuijlenborg	OGMION			n
B007QOS5T6	Axis Mundi	Cort Lindahl		y	2	y
B007QOSOG0	Dog Photography: A Point & Shoot Manual (Point & Shoot Manuals)	Jan Azana	Plutagora LLC	y	1	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QOVZES	Little Gems (12 brilliant short stories by a sick and perverted mind).	Ryan Hart		y	0	y
B007QOZ1AC	2600 Magazine: The Hacker Quarterly - Mac/PC - Spring 2012			n		n
B007QP4D6Y	Resonance & Vengeance (Combo Pack)	A.J. Scudiere	Griffyn Ink	y	3	y
B007QP4GMA	Through the Fire: Overcoming Trials, Tests, & Tribulations	Kim Harris		y	3	y
B007QP4IFU	Silvia's Abduction (Saorinan Chronicles)	Julie Sideris		y	1	y
B007QP5UQ6	An Annotated Bibliography of the US Marine Corps in the First World War	US Marine Corps		y		n
B007QP5V2Y	Caring for Roses - "Learn How to Grow Roses from Seeds, from Cuttings, in the Ground, in Containers... " Special Edition!	Michelle A. Rohn		y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QP5YVM	MONEY MANTRAS: THE POWER OF THE VEDIC HYMNS: MANTRA TO INVOKE THE GOD & GODDESS OF WEALTH & PROSPERITY	Kumar		y	0	n
B007QP7BUE	Awake Unto Me	Kathleen Knowles	Bold Strokes Books	n	1	y
B007QP8WYS	Quick & Easy Greek Yogurt Recipes: 47 Delicious "Almost Vegetarian" Greek Yogurt Dishes for Breakfast, Lunch, and Dinner (Quick & Easy Meatless Recipes)	Patty Douglass	J.J. Fast Publishing, LLC	y	0	y
B007QPBF0	Flying the Coop: The Video Game Mystery Novel	John Sailors	Story Crest Press	y	4	y
B007QPKTY4	Famous Shakespeare Sonnets	William Shakespeare		n		n
B007QPOXDW	Trucks! Big Trucks Doing Hard Work! (Over 25+ Photos of Awesome Trucks Working With Descriptions)	Cyndy Adamsen		y	0	y
B007QPRZ1O	The Twisted Thriller Files	Anna Katharine Green		n		n

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007QPU7D2	Fuck Scouts: Public Sex	Stroker Chase		y	2	y
B007QQ1BJA	Destiny Kills (6th Sam Casey Mystery)	S.D. Tooley	Full Moon Publishing LLC	y	3	y
B007QQ6MF8	98 Minutes to Paris	Marko Peric	Marko Peric	y	0	y
B007QQDTSQ	The Bread Man: A Kindness Never Forgotten	Will Bevis		y	3	y
B007QQR2M0	A Monster Got Me !	Charles Stippick		y	2	y
B007QQR39M	Shadow Demon (An Isaac Blackstone Novel)	Edward R. Murphy		y	4	y
B007QQUGPU	Little Citizens (Annotated)	Myra Kelly		n		n
B007QR52C6	Tess and the Star Traveller	Jane McKay	eText Press Publishing	y	3	y
B007QRCX14	One Life to Ride - A Motorcycle Journey to the High Himalayas	Ajit Harisinghani		n	2	y
B007QRIJQC	Personalentwicklung im Verein - Der Vorstand im Spannungsverhältnis zwischen Ehrenamt und Professionalität (German Edition)	Jessica Gießler	DVO Druck und Verlag Obermayer GmbH			n
B007QS0SGU	RUNZELWALD: Auf der Suche nach der kleinen Schwester	Adelheid Wildberger-Cattaneo	BookRix GmbH & Co. KG			n

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
	(German Edition)					
B007QT2BOG	Legends of Altai - Volume 1 - The Merchant Learns a Life Lessons	Paolo Tiberi		y	2	n
B007QTKCPQ	Fairy Night	Helen Ayim	JMS Books LLC	n	2	y
B007QTW6G4	Miss Leslie's lady's new receipt-book: a useful guide for large or small families containing directions for cooking, preserving, pickling	Eliza Leslie		n		n
B007QU2CO4	In Diplomatic Circles (Diplomacy)	Chris Tucker	Chris Tucker	y	1	y
B007QU2DSO	Sketches from a Celestial Sea - Gargoyle	Rob Heinze		y	2	y
B007QU2H30	Climate of Despair? The Future of U.S. Climate Policy and Global Negotiations	Nigel Purvis	German Marshall Fund	n	1	y
B007QUBSZI	Both Sides Of The Bars	Errol F Westfall	Errol F Westfall	y	0	y
B007QUDMHU	The Moving Picture Girls Or, First Appearances in Photo	Laura Lee Hope		n		n

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
	Dramas [Annotated]					
B007QUK69M	Sketches from a Celestial Sea - An Icy Christmas Candle	Rob Heinze		y	2	y
B007QUL6U0	Tess of the d'Urbervilles (Annotated) Characters Analysis, Themes, Motifs, Symbols & Study Questions	Thomas Hardy	OS Book	n		n
B007QUO9IQ	The Dog's Philosopher	Getrude Percy	eText Press Publishing	y	0	y
B007QUQT3E	THE LETTERS OF HENRY JAMES (Completed Volume I+II)	Henry James		n		n
B007QUU2HI	THE LIFE AND TIMES OF HINCMAR, ARCHBISHOP OF RHEIMS, A.D. 806-882	JAMES C. PRICHARD		n		n
B007QUW9A6	Hazard Ahead	Andy Hall		y	2	n
B007QUXG0I	Ravenswood and Custard	Mike Arkinstall	Olympia Publishers	y	0	y
B007QUZ9KI	Lost in Thailand	Jeffrey Johnson	booksmango	y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QUZRH8	Sketches from a Celestial Sea - A Day on the Minefields	Rob Heinze		y	2	y
B007QV0SIU	Cold	William J.Farrell	Trafford	y	4	y
B007QV3R7E	Northanger Abbey (Illustrated)	Jane Austen		n		n
B007QV43G8	The Middle Years (Annotated with Biography of the Author)	Henry James		n		n
B007QV47C8	Knowing God's Will	Peggy Billiard	Christianity Today	n	2	y
B007QVAQIW	The Goop Directory of Juvenile Offenders Famous for their Misdeeds and Serving as a Salutary Example for all Virtuous Children with Illustrated	GELETT BURGESS		n		n
B007QVARGI	Get Your Wings (Bonk! 3) (Bunk Ups!)	Jack McQueen		y	0	y
B007QVAWPY	Euro 2012 schedule. Group C. Essentials fan	Jarson24		y	0	y
B007QVBCZI	William Blake - Poezi dhe Poema (Albanian Edition)	William Blake	TOENA			n
B007QVE5ZM	Drive to Learn!	Srikanth Medi	Sri Medi	y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QVHVEO	Difficult Times - A Weekly Serial - Episode Twelve	Walter Shillington	Walter B Shillington	y	1	y
B007QVJUO8	THE LITERARY SENSE BY E. NESBIT (INCLUDING BIOLOGY OF AUTHOR)	E. NESBIT		n		n
B007QVJV0	Show Me the Trade: Revolutionary BLACK BOX for Profit in Stock and Options	Ron Groenke	Keller Publishing	y	3	y
B007QVJXHW	Sailing Across Europe	Andromeda Publications	Andromeda Publications	n		n
B007QVRXBK	The Watchtower Jesus	D. Allen Jenkins	Evangel Publishing House	n	2	y
B007QVRYI2	FISHING AND HUNTING : WITH PICTURES	SARAH M. MOTT	4U	n		n
B007QVU4W0	Having My Way With Words (From The 99 Part Of The Mind Poems Across Time)	Jay Olyan	BOBAIR MEDIA INC	y	1	y
B007QVUMW2	Diet Crock Pot Recipes (Easy)	Muffy Murphy		y	0	n
B007QW25O4	Okay For Now by Gary Schmidt I Summary & Study Guide	BookRags		n	2	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QW26YS	Comforted By A Cop	Arabella Keppler		y	1	y
B007QW2ADK	About Forex Trading - Buy Now	Simon Welch		y	0	y
B007QW2E5Y	The Simple A-Z Of Pregnancy - Tips & Tricks For Mothers And Mothers To Be.	Ray-Anne Blake	Ray-Anne Blake	y	2	y
B007QW4W4K	The Dream	Virginia Hansen		y	1	y
B007QW4Y8O	The True Gospel Message	Bonnie Lord	Bonnie Lord	y	0	y
B007QWEG4	James Baldwin - A Short Biography for Kids	Nell Madden	Shamrock Eden Publishing	n	0	y
B007QWEGM8	Mitchelhurst Place: A Novel, Volume 1	Margaret Veley		n		n
B007QWEH7W	Linux Journal April 2012		BELLTOWN MEDIA INC	n		n
B007QWEHWM	Mateja's Vegan Cookbook Collection	Mateja Tea Dereani	Fotospring E-Publishing Ltd	y	0	y
B007QWEPV0	Dream Shorts	Muriel Akamatsu		y	1	y
B007QWEQDC	Excerpts From The Alleys Of My Mind	John Lawrence Breska	John Lawrence Breska	y	1	y
B007QWSIXQ	The Discerners: Copperhead Road	Terry Halfhill		y	1	y
B007QWSNNQ	Martin	TR Montessor		y	0	y
B007QWXDHW	Deflowering Sir William	Victoria Scarlett		y	1	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QWY9WA	BEYOND the IVY LEAGUE: A Chosen Career Path	Pearl Chase		y	4	y
B007QXL2OC	American Legends: The Life of Ulysses S. Grant	Charles River Editors		n		n
B007QXQCG0	To Meat Or Not To Meat	Birgit Amadori		y	1	y
B007QXT7R6	EXTENDED EDITION You Should Never Go Home Again	Peter Butterworth	Peter Butterworth	y	2	y
B007QXWCFU	Draw Something Game: Play Online for Free, Get Instant Cheats, Tips, and Tricks	Peter Johnson		y	1	n
B007QY4R4S	Shawnee's Family Reunion Affair	Whiskey McNaughton	Crescent Suns eBooks	y	3	y
B007QY4SJM	I've Got A Choice	Dick Peterson	Juggernaut Press	y	0	y
B007QY4SZG	The Rampage	K.V. Black		y	0	y
B007QY4V7G	Clonuter	Monica Murray		y	1	y
B007QY61BU	Eine Liebe, die stotternd beginnt (German Edition)	Sylvia Seyboth	Sylvia Seyboth			n
B007QY62K0	Hannah - Das Ende vom Tag (German Edition)	Emma Janssen				n

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007QY63N6	Financial Investments, How & Where to Profit in Today's Market	Robert Manning		y	0	y
B007QY640S	3 Most Popular Ways To Take Vocal Lessons Reviewed: An honest look at the pros and cons of the 3 most popular ways people take vocal lessons	Ron Cross	<a href="http://www.TheMusicMinistryCoach.com">Http://www.TheMusicMinistryCoach.com</a>	y	2	y
B007QY78AI	Barbecue Secrets For Mouth Watering Favorite Foods!	Samuel Falope		y	0	y
B007QYC7HW	Escape to Texas	Kelly Volcik	DewClaw Productions	y	1	y
B007QYD3WK	The Complete Works of Mark Twain (Annotated)	Mark Twain	Gentoo Classics	n		n
B007QYD604	El Canto del Pelo de Hilo de Araña (Spanish Edition)	Julio Luis Ezpeleta				n
B007QYD80W	The Ugly Duckling (Illustrated)	Hans Christian Andersen		n		n
B007QYEQQ2	Sins of the Family	Adesokan Johnson	Adesokan Johnson	y	0	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007QYET1E	PROTEINAS G y sus Correlaciones GlicÃ³micas, ProteÃ³micas, MetabolÃ³micas y AntocianÃ³nicas en NanofemtofisiologÃ³a Vegetal para Agricultura Protegida (Spanish Edition)	Luis Alberto Lightbourn	Fabro Editores			n
B007QYJ9CS	Seven Important Things - Women Need from Men	Joseph Corrales	Joseph Corrales	y	0	y
B007QYPYJU	The Person Who Puts Down the Keys	DeAnna Knippling	Aurora Publishing LLC	y	2	y
B007QYVV5G	Merry's Marauders (Scenic Route to Paradise)	Andrea Aarons		y	1	y
B007QZ774Y	Lesser Gods	Tyler Vitt		y	2	y
B007QZB7BS	RehÃ©n de tu amor (Highlands) (Spanish Edition)	Brianna Callum				n
B007QZQM3G	Successful Camping for Everyone -- "Finally!The One You've Been Waiting For-AAA+++"	Tracey Warren		y		n

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007R07FZE	The Ultimate Yummy Low-Fat Cookbook Collection - Fast, Easy & Yummy Low-Fat Cookbook Collection, Vol. 1	Desiree Stewart	YumYum101, LLC	y	2	n
B007R07H02	Understanding and Applying the Bible	Dave DeLuca	Christianity Today	n	0	y
B007R07JN2	La GuÃa Definitiva Entrenar con Pesas para Correr (Spanish Edition)	Rob Price	Price World Publishing			n
B007R07P5O	The End of a Rotten Day	Dave Hendrickson	Pentucket Publishing	y	3	y
B007R0BGNQ	Striped Stocking Cap Hat Tam Beanie & Gloves Knit Knitting Pattern	The Crochet Kid	The Crochet Kid	y	0	y
B007R0HFC2	Naked God	Martin Ayers	Matthias Media	n	1	y
B007R0HH8O	Complete Defense to Queen Pawn Openings	Eric Schiller	Cardoza Publishing	n	3	y
B007R0IPKI	Wannabe Spy Club	A.J. Clare	Barton Publishing	y	4	y
B007R0IQUU	WRITERS' AUTHORITY	SAM HUNSU	PERFECT PROMISE COMMUNICATIONS	y	1	n
B007R0KZK6	Lust In Space	Robert Fittro	Pushbutton Studios	y	0	y
B007R0KZTC	Overtime	Robert Fittro	Rooster Magazine	n	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007R0NN46	English-Spanish Legal Glossary (Learn Legal Terms Translated into Spanish)	Esperanza Lopez		y	0	y
B007R0NNH8	Susan's Leaf Crochet Pattern Collection	Susan Kennedy		y	1	y
B007R0P862	The Spoils of Poynton (Annotated with Biography of the Author)	Henry James		n		n
B007R0P8EE	Beautiful Wild Rose Girl	B. Magnolia	Mystic World Press	y	0	y
B007R0RQ3A	Dismantling the Twin Towers of Race and Racism: A Vision of America's Way Forward	Alfred M Walker I	Searchlight Press	n	0	y
B007R0RQR6	Siren Island: Lust For Gold (An Erotic Adventure Series)	Virginia Wade	I Love Stacy	y	4	y
B007R19S12	Pronouncing the Mackay-Bennett	Daniel Eness	Eortholic Press	y	3	y
B007R1DK3Y	Thamos, King of Egypt	Tobias von Gebler	Createspace	y		n
B007R1WASK	Savage (New Atlantis)	Nhys Glover		y	0	y
B007R2HZYS	Mfuko Wangu U Wapi?	Akberali Manji	Phoenix Publishers and Worldreader		0	y
B007R2KRPW	Oma	Lily Mabura	Phoenix Publishers and Worldreader	n	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007R2ODD4	Slovoed Deluxe French-Russian dictionary (Slovoed dictionaries) (French Edition)	Ruskiy yazik Media				n
B007R2OE5G	Beyond Silence	Kusum Ansal	Roli Books	n	0	y
B007R2TQCM	Slovoed Deluxe German-Russian dictionary (Slovoed dictionaries) (German Edition)	Ruskiy yazik Media				n
B007R2TWL2	Siri ya Sala (Swahili Edition)	Njiru Kimunyi	Phoenix Publishers and Worldreader			n
B007R2Z1JE	Silver and The Secret of Troika	Blake Collins	Blake A. Collins	y	0	y
B007R321KK	Slovoed Deluxe Catalan-Spanish dictionary (Slovoed dictionaries) (Catalan Edition)	S.L. Larousse Editorial				n
B007R374SE	Das Leben und Sterben des Maximilian (German Edition)	Ham Pash Chadaev				n
B007R3AG0W	How To Become An Internet Traffic Broker	DONATUS AMAECHI		y	1	n
B007R3JISI	Slovoed Classic Uzbek-English dictionary (Slovoed	Ulugbek Isakov		y	0	n

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
	dictionaries)					
B007R3QW8C	Works of Dr. John Tillotson, Late Archbishop of Canterbury. Vol. 8	John Tillotson		n		n
B007R3R11E	HOW TO BUY A HOME USING A VA LOAN: What Every Home Buyer Should Know	Stacey Chillemi	Lulu Inc.	y	3	y
B007R3SXM0	Woody Allen Biografía (Spanish Edition)	Adolfo Pérez	Ediciones Masters			n
B007R3Y1EO	Drive Business Growth with the Telephone	Valerie Schlitt		y	3	y
B007R40RUU	Newton's Principia:[Illustrated]	SIR ISAAC NEWTON		n		n
B007R40UO8	The Autism Handbook: Easy to Understand Information, Insight, Perspectives and Case Studies from a Special Education Teacher (Simplified Chinese Edition)	Jack E. George	CCB Publishing	y	3	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007R42F6Y	The Art Of Kissing	Hugh Morris		y	1	y
B007R4CLBI	In Satans Namen (German Edition)	Andreas Schmidt	Rhein-Mosel-Verlag			n
B007R4CS8Y	Love, Light, and Joy: A 90 Day Practice That Will Change Your Life	Brian Reekers	Angelic Intuition	y	3	y
B007R4IAN6	HORACE CHASE. (ANNOTATED)	CONSTANCE FENIMORE WOOLSON	Amazon.com	n		n
B007R4IB5I	TONIO, SON OF THE SIERRAS A STORY OF THE APACHE WAR BY THE AUTHOR OF MARION'S FAITH, A DAUGHTER OF THE SIOUX, MARION'S FAITH : AMERICAN WESTERN CLASSIC STORY (Annotated)	Charles King	timelessbook.info			n
B007R4IBDA	Ogham Symbols Aura Metalexicon Logodynamics	Gregory Zorzos		y	1	y
B007R4M8K2	Alexa Sees Farm Animals (Personalized Book with the name Alexa)	Mike Fawn		y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007R4U9VW	Alexandra Sees Farm Animals (Personalized Book with the name Alexandra)	Mike Fawn		y	0	y
B007R4WK6Y	Mabel: A Novel, In Three Volumes (Volume 3)	C. J. Mrs Newby		n		n
B007R4WKLY	The Book of Jubilees	R. H. Charles		n		n
B007R50PX8	Tax-Free Wealth	Tom Wheelwright	BZK Press	y	5	y
B007R56Q5O	Steven Moffat's Doctor Who 2011: The Critical Fan's Guide to Matt Smith's Second Series (Unauthorized)	Steven Cooper	Punked Books	n	2	y
B007R59PRA	El paso de los espa±oles (Spanish Edition)	VerÁ³nica MartÁnez Amat				n
B007R59W32	Thumbelina (Illustrated) (Andersen's fairy tales)	Hans Christian Andersen	The Planet	n		n
B007R5EO8K	Ariana Sees Farm Animals (Personalized Book with the name Ariana)	Mike Fawn		y	0	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007R5J67O	THE TRAIL HORDE BY CHARLES ALDEN SELTZER AUTHOR OF THE BOSS OF THE LAZY Y, THE TRAIL HORDE, THE RANCHMAN : AMERICAN WESTERN CLASSIC STORY (Annotated)	Charles Alden Seltzer	timelessbook.info			n
B007R5J76Y	Beyond Human - Alien and Tentacle Sex Erotica Bundle	Katie Cramer	Addictive Press	y	4	y
B007R5J8HW	When her Ex became her Boss	L Lovett		y	0	y
B007R5J98K	The Virgin's First Time (contemporary romance, erotic romance)	D K		y	0	y
B007R5JA9S	Me, Myself and I (Dark Reflections)	Natasha Duncan-Drake	Wittegen Press	y	4	y
B007R5JETE	Accidental Alphabet (Nonsense Verse Books)	Charmian Hayes		y	2	y
B007R5LNP2	Jonah	James D. Quiggle		y	3	y
B007R5Q2Q2	Tor (WereWolf Fight League)	Lynn Lorenz	Loose Id LLC	n	4	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007R5TCWI	Dictionary of the Hausa language (Volume 1)	Charles Henry Robinson		n		n
B007R5TEDA	As A Men Thinketh	James Allen	Scapen Books	n		n
B007R5TIEU	Dramatically Raising Your SAT score	Darryl Hold	Darryl Hold	y	0	y
B007R5TJ8U	The Legend in the Cloud	A. J. McGane	A. J. McGane	y	1	y
B007R5VMQM	Les Etonnants Discours d'Investiture des PrÃ©sidents de la RÃ©publique (French Edition)	Franck de MAGALHAES				n
B007R5WSX8	Eye of the Beholder	J Charnock-Shields		y	1	y
B007R5WV08	Family Stories (True Funny and Scary Family Stories)	Chappell Mosley		y	0	n
B007R5X01W	The Gamma Collection, Volume 4	Samir Patel		y	0	y
B007R5ZWJ0	FDA: NDA and IND Applications - Contemporary Decisions (Health Care Law Series)	LandMark Publications	LandMark Publications	n		n
B007R600Q4	Forbidden Love (Erotic Romance)	Ella Fisher	Ella Fisher	y	0	y
B007R60NTS	THREE HILLS (The Borderer Chronicles)	Mark Montgomery		y	4	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007R650UA	ARSÃ^NE LUPIN - Le Bouchon de cristal (ARSÃ^NE LUPIN GENTLEMAN- CAMBRIOLEUR) (French Edition)	Maurice Leblanc	Collection Lupin			n
B007R65262	ARSÃ^NE LUPIN - La Comtesse de Cagliostro (ARSÃ^NE LUPIN GENTLEMAN- CAMBRIOLEUR) (French Edition)	Maurice Leblanc	Collection Lupin			n
B007R65348	ARSÃ^NE LUPIN - La Cagliostro se venge (ARSÃ^NE LUPIN GENTLEMAN- CAMBRIOLEUR) (French Edition)	Maurice Leblanc	Collection Lupin			n
B007R654GK	ARSÃ^NE LUPIN - La demeure mystÃ©rieuse (ARSÃ^NE LUPIN GENTLEMAN- CAMBRIOLEUR) (French Edition)	Maurice Leblanc	Collection Lupin			n
B007R6D2FA	Oberon (formerly published as Homecoming)	Barbara Bickmore	Barbara Bickmore	y	4	y
B007R6FCVW	For the Sake of	Andrew Hill		y	1	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
	Survival					
B007R6FD20	Steam: a rock and rail tale	Walt Oxley		y	0	y
B007R6HX3C	Getting Your Best Health Care: Real-World Stories for Patient Empowerment	Ken Farbstein		y	0	y
B007R6ILU6	Paralyzed (A novel.)	John Meany		y	3	y
B007R6KYRO	Mazurek ground-cranberry aroniowo. Polish Cake. Polish Kitchen	Jarson24		y	0	y
B007R6M7KQ	Starvelings	S.D. Hintz	Aristotle Books	y	5	y
B007R6OAOA	10 Website Traffic Tips	anik		y	0	y
B007R6OOFC	Route To Peace 2	Fidel Nshombo	Borderline Publishing	y	4	y
B007R6OOH0	Poison mysteries in history, romance and crime	C.J.S. Thompson		n		n
B007R6OPV0	Sex,Lies and Love (1)	Alice Simpson		y	1	n
B007R6S0PW	LA FÃ‰ODALITÃ‰ CHINOISE (French Edition)	Marcel GRANET				n
B007R6S1H4	Dictionary of the Hausa language (Volume 2)	Charles Henry		n		n
B007R6S5CU	What I Did On My Midlife Crisis	Debbianne DeRose		y	5	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
	Vacation					
B007R6TE66	MELANGES sur L'ADMINISTRATIO N (French Edition)	Pierre HOANG				n
B007R6VX60	Human Development Report 2006: Beyond Scarcity - Power, Poverty and the Global Water Crises	United Nations Development Programme (UNDP)	United Nations	n		n
B007R71ZPI	La vita comincia domani: romanzo (Italian Edition)	Guido da Verona				n
B007R71ZQW	La Segunda Guerra Mundial, Europa (Spanish Edition)	J. G. Berenguer				n
B007R722DW	Segreti Automatici - Seconda Raccolta (Italian Edition)	Claudio Facilla				n
B007R72Z48	Canti popolari raccolti in Napoli. Con varianti e confronti nei vari dialetti (Italian Edition)	Molinaro del Chiaro				n
B007R7322W	Les Tarots du Chat in 22 Arcani Maggiori (Italian Edition)	Evelyne Nicod	Gatteria			n

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007R77E5S	SharePoint 2010 Issue Tracking System Design, Create, and Manage	Sarath Thirumoorthi		y	0	y
B007R79PIC	General Book of the Tarot	A. E. Thierens	Mariana de Lacerda Oliveira	n		n
B007R79QCC	Killing In 3's, Tales of Murder and Mystery	William Makepeace Thackeray		n		n
B007R7B4YA	L'arc en ciel ou les couleurs de notre temps (French Edition)	Nadia Rafin	La Plume de l'ArgilÃ´te			n
B007R7B832	Autism Spectrum Disorders: A Parent's Guide to Autism and Asperger Syndrome	Sarah Erickson		y	0	y
B007R7B8CS	Honest Insincerity (A Variety of Passion)	James Baumann	Postmortem Publications Inc	y	0	y
B007R7FA40	Trayvon Martin: Injustice In America	Eric Benoit		y	3	n
B007R7FAU4	The Artist Charge (A Variety of Passion)	James Baumann	Postmortem Publications Inc	y	0	y
B007R7XBWI	Army Field Manual - Mortar Fire Direction Procudures	US Army		y		n
B007R8B5N4	The Fund	Jeff Edwards	Port Campbell Press	n	0	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007R8BFHU	THE NAUGHTY PLEASURES BUNDLE (Naughty Pleasures: Volume 1/Naughty Pleasures: Volume 2/Naughty Pleasures: Volume 3)	Abbie Cole		y	2	y
B007R8H55G	Frenzied Feast of the Maenad	Mallorie Griffin		y	2	y
B007R8H6YQ	Repeal of "Don't Ask - Don't Tell "	Department of Defense		y	0	y
B007R8LYMG	The Logs of the Watersnake and Water Rat	H. Fiennes Speed		n		n
B007R9AE62	Pearls of Wisdom - Pure & Powerful	Dr. Liz Anderson-Peacock	International Health Publishing	y	4	y
B007R9G1F0	Oeuvres d'Alexandre Dumas (Illustr�e) (French Edition)	ALEXANDRE DUMAS	Delphi Classics			n
B007R9G6GE	Doctor's Orders 3	C.M. Knox		y	0	y
B007RBA5IW	Vegetarian Cooking: Root Vegetables Soup (Vegetarian Cooking - Soups)	Wancy Ganst	Wancy Ganst	y	2	y
B007RBADMK	Secondary School 'KS3 (Key Stage 3) Maths - Geometry and Measures - Ages 11-14' eBook.	John Kelliher		y	3	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RBF4L0	Vegetarian Cooking: Marinated Gluten, Radish and Zucchini (Vegetarian Cooking - Vegetables and Fruits)	Wancy Ganst	Wancy Ganst	y	2	y
B007RBMNHI	The Warriors: The Dance, Dinner and Danger	Marvin Rondares		y	2	y
B007RBS66U	20 Top Fitness and Health Tips for Busy People: quick and easy health, fitness and diet tips for men and women that can promote healthy living, improve fitness, aid weight loss and maintain motivation	Darren Evans	Darren Evans	y	0	y
B007RC2ZA2	Back Pain Busters : Home Remedies and Natural Cures for Back Pain (Back Pain Relief)	Herbert Webster	NTC	n	1	y
B007RD8Z6Y	Un momento de inspiraci3n (Spanish Edition)	Vicens Jordana	edicions2.0			n

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007RDBUII	Les tribulations d'un chinois en Chine (Annoté par Lycium Classiques) (Voyages Extraordinaires) (French Edition)	Jules Verne	Lycium Classiques			n
B007RDKFAI	Goddess of Thunder	Lilah Wild	Leopard Moon Press	y	0	y
B007RDZ1J8	Grand Slam: A Tale Of Weird Golf	Dave Donelson	Donelson SDA, Inc.	y	3	y
B007REEMD8	Swann's Way (Annotated) Characters Description, Themes & Study Questions	Marcel Proust	OS Book	n		n
B007REIH XO	Chinese-Japanese Cook Book	ONOTO WATANNA		n		n
B007REII8I	Middlemarch - [ Free Audiobook Download ] [ Annotated ]	George Eliot		n		n
B007REJ2Q0	Love is the Reason	Mary Malone	Poolbeg Press	n	5	y
B007RESZZ4	Campaigns of World War II: A World War II Commemorative Series - Ryukyus	Jr. Arnold G. Fisch		y	0	y
B007RET2JM	BLACK BEAUTY - [ Free Audiobook Download ] [ Annotated ]	Anna Sewell		n		n
B007RET3YQ	These Three	John DiLeo	Hansen Publishing Group	n	3	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RET7IS	Zombies! Episode 2.6: The Many Deaths of John Arrick	Ivan Turner		y	3	y
B007RF1HU8	Um Vagabundo em New York (Portuguese Edition)	Chafi Nader				n
B007RF1ISO	Candy Makes Three	Joey Dueck		y	3	y
B007RF1KRS	La niÑ±a del mar (Spanish Edition)	RamÑ±n VillerÑ±				n
B007RF2AWM	Astronomical Curiosities (Illustrated)	J. Ellard Gore		n		n
B007RF2D24	The Divine Law . . . Love Awareness Wisdom . . . a spiritual journey	Allan Smith		y	0	n
B007RF8ATE	THE PROFESSIONAL-REVISED EDITION	RICHARD GEARY		y	1	y
B007RF8FEY	Make Mine A Double (Lacey's Lamp)	Elizabeth Coldwell	MuseItHOT Publishing	n	2	y
B007RFASY4	Better Sex without Viagra (Assistant Mistress)	Teri Power		y	1	y
B007RFG12M	The Imperial gazetteer of India	Great Britain India Office		n		n
B007RFG1B8	A Winter Nosegay: Being Tales for Children at	Walter Crane		n		n

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
	Christmastide					
B007RFG2AS	Truth? Or lies from the pulput?	Simple Carpenter		y	0	y
B007RFI7CE	Fly Away Home	Maggie Myklebust	Summertime Publishing	n	4	y
B007RFIMGA	The As You Wish Series, Volumes 1-3	Mindy Klasky	Res Ipsa Press	y	3	y
B007RFIMQA	Bumping Off Binky (A Curl Up and Dye Mystery - Book 2) (The Curl Up and Dye Mysteries)	Nancy Mehl	Greenbrier Book Company	n	3	y
B007RFOVVK	Protecting Miss Samuels	Summer Devon		y	6	y
B007RFP05G	Mail Order Mistake?	Mary C. Findley	Findley Family Video	y	3	y
B007RFTF84	Personify This	Donald Enz		y	0	y
B007RFTFSY	The Reality of Life	Hans-Juergen Briest	Hans-Juergen Briest	y	1	y
B007RG6BAS	Nominigan and Other Smoke Lake Jewels	Gaye I. Clemson	FastPencil, Inc.	y	1	y
B007RGGBRG	Drinking Lydia's Come (Daddy's Tails)	Tasia Winters		y	4	y
B007RGGBDGK	Sunnyvale	Hellen Fellowes	Academy Incorporated Ltd.	y	0	y
B007RGBNDI	Beyond the Philadelphia Experiment and Montauk Project	Bill Nelle		y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RGCVAC	On Piratical Seas: A Merchants Voyage to the West Indies in 1805	Peter Grotjan	Folly Cove 01930	n		n
B007RGPGJU	The 21st Century Toolbox: Innovative Apps for Educators and Students	Monica Sevilla		y	2	y
B007RGPLX6	Portais do CÃ©rebro (Portuguese Edition)	Cleser Campos	Cleser Campos			n
B007RGPOU6	Human Development Report 2003: Millennium Development Goals - A Compact Among Nations to End Human Poverty	United Nations Development Programme (UNDP)	United Nations	n		n
B007RGPOWE	Human Development Report 2010: The Real Wealth of Nations - Pathways to Human Development	United Nations Development Programme (UNDP)	United Nations	n		n
B007RGT22C	The Chronicles of Eleanor Rose - The Sunnyville Chapter	Paul Laurent		y	1	y
B007RGYVLE	U. S. A. THE CORPORATE STATE	LES CRANE		y	0	y
B007RH0B1M	After Vietnam: Paths to Love and Learning	Gabriel Urbina	Gabriel Urbina	y	1	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RH5PF4	Dark Dates (Cassandra Bick Chronicles)	Tracey Sinclair		y	3	y
B007RH5VYO	The Violets Are Mine	Lester Morris	Javelina Books	y	2	y
B007RH72CS	Clone Whores	Saffron Smith		y	2	y
B007RH8178	GUM - (Children's Picture eBook)	Lydia Scherr		y	0	y
B007RH83QC	A Beast in the Sack: My Boyfriend the Werewolf	Wynne Burroughs		y	0	y
B007RHLVPW	Tales of Tomorrow Girl: Retro Science Fiction Adventures of Mystery and Romance	Robert Szeles		y	6	y
B007RHW9IU	God Is My Teddy Bear (Religion--Who needs it? And Why.)	Ted Farris	TN Farris	y	0	n
B007RHWBB0	Creative Gratitudes Vol 2	Felipe Adan Lerma	Yoga-Adan eBooks	y	3	y
B007RHXV8M	Game of Life Down Under	Kazmax	Kazmax	y	0	y
B007RICJMK	Moving Forward	Steven Reid Swanson	Steven Reid Swanson	y	0	y
B007RICNFS	Le renouveau de la terre (French Edition)	ComitÃ© Pre Ohm	Osmora Publishing			n
B007RIG8T0	The Atlantis Precedent	David McGarry		y	1	y
B007RIHYXE	ENUFF: Eliminate the Needless, Useless, Foolish, and Frivolous	Kate Carpenter		y	3	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RJFY5I	The Never Men (Tales of Tomorrow Girl)	Robert Szeles		y	6	y
B007RJOA9E	The Alchemists' Circle (The Adventures of Bouragner Felpz)	Goldeen Ogawa	Goldeen Ogawa	y	3	y
B007RJOEJO	Our Favorite Place	Mariah Walker		y	2	y
B007RKKYRK	The Naval Pioneers of Australia (Illustrated)	Louise Becke	DCR Publishing	n		n
B007RKQ95G	Eternal Youth	Julia Crane	Valknut Press	y	5	y
B007RKRHGG	Sailing with Senta - Tropical Dream	Faith and Pierre Van Rooyen		y	3	y
B007RL29WM	Welsh Executions	John Eddleston	Bibliofile Publishers	y	2	y
B007RL9ARO	Controindicazioni. Perch� il rock fa male (Italian Edition)	Tommaso Franci				n
B007RMF6W6	Out of the Shadows (The Order)	L.K. Below	Lyrical Press, Inc.	n	3	y
B007RMP9XC	The smoke of the fatherland	Joseph Nemlicher				n
B007RMTOUG	Berserk Revenge	Mark Coakley		y	1	n
B007RMWMIC	When Jesus Came to Jersey as the Son of Thunder (Aaron Adams Adventures)	Wayne Frye	Peninsula Publishing	y	0	y
B007RMXLQE	The Picture of Dorian Gray By Oscar Wilde (Oscar Wilde Version and Illustrated)	Oscar Wilde		n		n

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RN03ZA	O + E Words ("Magic E") - Colour (Phonic Words Colour)	Billy Bumble	JayCurt LLP	y	4	y
B007RN2XTO	Murder Down Under	Nancy Curteman	Solstice Publishing	n	6	y
B007RN9NY2	Mildred Arkell: A Novel, Volume 2 (of 3)	Mrs. Henry Wood		n		n
B007RNJPOA	The Head (Pictures of Veterinary Anatomy)	Clemens Knospe		y	1	y
B007RNJRR0	Ultimate Diving Adventures	Bob Walker		y	0	y
B007RNJW7A	Captain Scott's Last Ten Days	Richard Pierce	Gerald Duckworth	n	4	y
B007RNMM90	Met Art Nude Erotica Model Sexy Hot v3	LADYHOT	LADYHOT	y	0	n
B007RNREY8	Exercise The Key to Good life	Johnson JC		y	3	n
B007RNRGQE	25 Fast & Simple Recipes for Chicken Authentic Mexican Dishes (Fast & Simple Chicken Recipes Cookbook Collection)	Gavin Ray		y	0	y
B007RNRL96	Amy Moaned	Lord Koga	Veenstra Publishing	y	3	y
B007RO0U84	ABC's for Achieving a Winner's Mindset (The ABC's for Succeeding In Life)	S. Williams		y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RO14SY	The Man In the Top Hat (Ghost Story)	John Meany		y	3	y
B007RO16EQ	Der Zapfhahn des Tankwarts (Bronco Baxter - Gay Story 1) (German Edition)	Tom Dillinger				n
B007RO8OFA	THE CURSE OF CARNES HOLD	Unknow	Unknow	n		n
B007RO8RRU	The Empty House and Other Ghost Stories (Fully Illustrated)(Annotated)	Algernon Henry Blackwood		n		n
B007RO919I	Sunstone 166 (Sunstone Magazine)	Tresa Edmunds	Sunstone Education Foundation Inc.	n		n
B007ROBQ7I	Labor Management Relations Act: Contemporary Decisions (Employment Law Series)	LandMark Publications	LandMark Publications	n		n
B007ROGQ8M	Hollow Earth	Leonard Euler		n		n
B007RON3X8	LA CHINE et la FORMATION de L'ESPRIT PHILOSOPHIQUE en FRANCE (1640-1740) (French Edition)	Virgile PINOT				n

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RON4R8	The Best of Lucy Maud Montgomery: Anne of Green Gables and Other Works (Annotated)	Lucy Maud Montgomery	Di Lernia Publishers	n		n
B007RONLVW	LIBBY, Proof there is a Supernatural Realm	David Hamilton		y	0	y
B007ROROH4	Hypertension Control And Cure Without Medication: Experience and Testimony of an Hypertensive Elderly Man .	Arthur Davis		y	1	y
B007ROVG7S	Hero Cat	Eileen Spinelli	Amazon Children's Publishing	n	3	y
B007RP2Y6O	Vipassana Meditation: My Experiences at a 10-Day Retreat	Robert Crayola		y	0	n
B007RP4ZCK	Building The BEST business team	s. blake		y	0	y
B007RP9TTY	The Positive Effects of Insomnia on the Human Mind	Mike Richardson		y	1	y
B007RPMFN6	My Thoughts, Exactly!	Jim Ross	Merlinseyes Press	y	2	y
B007RPRFUO	Karl Marx: Das Kapital	Karl Marx		n		n
B007RPTQTC	911 Finding the Truth	Andrew Johnson		y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RPTSC2	Promoting Mission Success for the USMC Distributed Operations Squad Through Efficient Equipment Selection	Shawn M. Charchan		y	0	y
B007RPTSYA	Evaluation of Organizational Self-assessment Tools and Methodologies to Measure Continuous Process Improvement for the Naval Aviation Enterprise	Theodore J. Kaehler		y	0	y
B007RPTTCQ	An Alien Collective	Roxanne Barbour	Fantasy Island Book Publishing	n	3	n
B007RPTVD8	Enabling System Management through Process Modeling: The Australian Defence Force Recruiting System	Carissa C. Ibbott		y	0	y
B007RPWEVE	Impact of Radio Frequency Identification (RFID) on the Marine Corps' Supply Process	Melissa D. Chestnut		y	0	y
B007RPWP04	You Can't Pick Up Raindrops	John Charles Miller	John Charles Miller	y	3	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RPWVNK	The Message of Brazilian Rituals (Brazil: Body & Soul)	Roberto DaMatta	Guggenheim Museum	n	2	y
B007RQ2ZSK	The Last English Village	James Ignizio	Cowan Creek Press	y	2	y
B007RQ563Q	Shopping for Ghosts (Shopping for Ghosts and Other Unusualities)	Adam Betley	Adam Betley	y	0	y
B007RQIV24	Falta el aire (Spanish Edition)	Antonio GÃ;lvez Alcaide				n
B007RQIWXC	Between "I Will" & "I Do"	James Olson		y	0	y
B007RQMIVY	The Girl Upstairs	Ray Sostre	Romance Divine LLC	n	5	y
B007RQM7A	Condemned by a Vampire	Jodie Pierce	Eternal Press (an imprint of Damnation Books LLC)	n	2	y
B007RQMS3W	Los lÃ;mites del mundo (Spanish Edition)	Carlos Almira	Editorial Amarante			n
B007RQOQ46	Unlock Blackberry all Model (Spanish Edition)	JGT GT				n
B007RQP3DO	The Haunted House	JJ Argus		y	4	y
B007RQP47O	The Alchemist Lesson Plans	BookRags		n	2	y
B007RQV0BS	PRATIQUE des EXAMENS LITTÃ;RAIRES en Chine (French Edition)	P. Ã;otienne ZI				n

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007RQV6KI	Fallout	Karlene Blakemore-Mowle	Eternal Press (an imprint of Damnation Books LLC)	n	4	y
B007RQVDAG	Filariasis: Causes, Tests, and Treatment Options	James Garner MA		y	0	y
B007RQWG0W	Waiting For My Eyes to Adjust	J. Calvin Westbourne	J. Calvin Westbourne	y	0	y
B007RQWGAW	Les fluctuations Économiques À longue période et la crise mondiale (French Edition)	François Simiand				n
B007RQWW6K	Dorm Room Double Shots: After Curfew Ass Play & Lights Out, Cocks In	Julieta Hyde	Rutting Good Press	y	4	y
B007RQZ7NA	Vingt mille lieues sous les mers (Annoté par Lycium Classiques) (Voyages Extraordinaires) (French Edition)	Jules Verne	Lycium Classiques			n
B007RRDN5S	Your Guide to Congaree National Park	Michael Oswald	Stone Road Press	y	4	y
B007RRE0P0	Your Guide to Isle Royale National Park	Michael Oswald	Stone Road Press	y	4	y

Amazon ASIN	Title	Author	Publisher	Self Published?	Social Media Use Count <sup>24</sup>	Used in Regression ?
B007RRPYKA	The Pilgrim cook book : containing nearly 700 carefully tested recipes	Pilgrim Evangelical Lutheran Church		n		n
B007RS202E	The Gods of the Lodge	Reginald Haupt Jr	RiverCrest Publishing	n	0	y
B007RS8IV6	Amish Forever- Volume 7- Where Has Love Gone?	Crystal Linn	Trestle Press	n	1	y
B007RS8NMK	Say What?	Stephen Smith	Wild Publishing	y	3	y
B007RSBWPO	Freeze Frame	Kimberly Duquette	Kimberly Duquette	y	5	y
B007RSC7D6	A Death In Munich	David L. Hoof	Trestle Press	n	3	y
B007RSLNFO	The Chinese in Indonesia	Pramoedya Ananta Toer		n		n
B007RTD5HM	Trait d'Économie politique (French Edition)	Jean-Baptiste SAY				n
B007RTD7CK	Headquarters Department of the Army FM Crew Served Machine Guns, 5.56-mm and 7.62-mm	US Army		y		n
B007RTDDGK	The Unlimited	Kevin Basil	Kevin Basil	y	3	y
B007RTNZS6	Escaping to your Arms	Mary Heathcliff		y	5	y
B007RTO0EO	Mere Enchantment (The Enchantment Series)	Alicia Rivoli	Alicia Rivoli	y	3	y
B007RTX37O	The Hunger Fire	Janvier Chando	TISI BOOKS	y	2	y
B007RUBRLW	The Sword and the Storm	Jon-Eric Nissen		y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RUNEUO	Horror Vacui	Amanda Cale		y	2	y
B007RUTO0S	Lecture Sialkot	Hazrat Mirza Ghulam Ahmad of Qadian, India. Promised Messiah and the Guide		n		n
B007RV7GAW	Ignition (Wolf Pack Columbus)	Christian Kaylor		y	0	y
B007RV7K5S	Five Hands Diet	Emily Callowdic		y	0	y
B007RVD96M	What Is Tourism and Top 20 Best Places to Visit All Over the World	Venkataramana Rolla		y	0	y
B007RVDEPS	Cherry Addition	Joy Findlay		y	5	y
B007RVGQVW	Breaking the Thought Barrier	Rob Riser		y	0	y
B007RVWA44	Beat Depression Naturally (Dr. Singh's Guide)	Dr. Anurag Singh	HealthEnclave	y	6	y
B007RWGYHM	Tickle Your Amygdala	Neil Slade	Neil Slade Brain Books, Music, and Film	y	4	y
B007RWUUU4	The One Hundred Calorie Diet and Food Counter	Susie Trimble	The Wellness Institute Publications of America, LLC	y	2	y
B007RWY0VO	The Greatest Relationship Secret	Astra Niedra	Astra Niedra	y	5	y
B007RX090E	How to Build an Engine Hoist	Ian Toms	Bermuda Publishing	y	0	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published?</b>	<b>Social Media Use Count<sup>24</sup></b>	<b>Used in Regression ?</b>
B007RXCS DK	CHASING CORVINI	Tom Hurst	The Electronic Book Company	y	0	y
B007RXN9WY	Between chaos and consciousness - hyperphysics	Joanna Rajska	Jacek Czapiewski	y	0	y
B007RXPYFO	AD ASTRA 020 Heftausgabe: Die Marsel fe (AD ASTRA Heftausgabe) (German Edition)	K. H. Reeg	www.HARY-PRODUCTION.de			n
B007RXPZZS	THE VALUE OF THE SESTERTII (THE VALUE OF THE COINS)	Ernesto G.Guinea	Administraciondigital S.L	n	0	y
B007RXVBH4	Running with the Devil	Robert Long		y	1	y
B007RXVCQ4	The Royal Captive: Vol 1 to 3	Aphrodite Hunt		y	4	y

### APPENDIX C – POPULAR SAMPLE TITLES

Amazon ASIN	Title	Author	Publisher	Self Published ?	Social Media Use Count <sup>25</sup>	Used in Regression?
B000FBFN1U	A Storm of Swords: A Song of Ice and Fire: Book Three	George R.R. Martin	Bantam	n	6	y
B000FC1HBY	A Clash of Kings: A Song of Ice and Fire: Book Two	George R.R. Martin	Bantam	n	6	y
B000FCKGPC	A Feast for Crows: A Song of Ice and Fire: Book Four (Martin, George Rr)	George R.R. Martin	Bantam	n	6	y
B000JMLFLW	Pride and Prejudice	Jane Austen	Public Domain Books	n		n
B000JMLNHI	Fairy Tales Every Child Should Know	Hamilton Wright Mabie	Public Domain Books	n		n

---

<sup>25</sup> Number of Amazon page, Goodreads page, Facebook page, Web page, Blog, Twitter account

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B000JQU1VS	The Adventures of Sherlock Holmes	Sir Arthur Conan Doyle	Public Domain Books	n		n
B000JQV3QA	Alice's Adventures in Wonderland	Lewis Carroll	Public Domain Books	n		n
B000P28WPI	25 Days to Better Thinking and Better Living: A Guide for Improving Every Aspect of Your Life	Richard W. Paul	FT Press	n	3	y
B000QCS8TW	A Game of Thrones: A Song of Ice and Fire: Book One	George R.R. Martin	Bantam	n	6	y
B0015DROBO	The Girl with the Dragon Tattoo: Book 1 of the Millennium Trilogy	Stieg Larsson	Vintage	n	4	y
B0018QQQK8	The Lucky One	Nicholas Sparks	Grand Central Publishing	n	5	y
B001C34D0M	Heaven for Kids	Randy Alcorn	Tyndale Kids	n	6	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B001EOCFU4	The Holy Bible, English Standard Version (with Cross-References)	Crossway Bibles	Crossway	n		n
B001JEPF4W	Murder on the Rocks (Gray Whale Inn Mysteries, No. 1) (Gray Whale Inn Mystery)	Karen MacInerney	Midnight Ink	n	6	y
B001LDJQM	The Apothecary's Daughter	Julie Klassen	Bethany House	n	4	y
B001N2MBJW	Paper Roses (Texas Dreams Trilogy #1)	Amanda Cabot	Revell	n	5	y
B001NLKT60	The Girl Who Played with Fire (Millennium Trilogy, Book 2)	Stieg Larsson	Vintage	n	4	y
B00280859I	The Crossroads Cafe	Deborah Smith	BelleBooks	n	4	y
B002E9IPSO	Blackbird Fly	Lise McClendon	Thalia Press	y	6	y
B002L4QP0M	A Matter of Honor	Jeffrey Archer	St. Martin's Paperbacks	n	6	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B002MQYOFW	The Hunger Games	Suzanne Collins	Scholastic Paperbacks	n	4	y
B002Y5W9NK	Elisha's Bones	Don Hoese	Bethany House Publishers	n	5	y
B002YK0XB6	The Help: Movie Tie-In	Kathryn Stockett	Berkley	n	5	y
B002YT8JD0	Jenna's Cowboy: A Novel (The Callahans of Texas)	Sharon Gillenwater	Revell	n	4	y
B002Z7G0QO	HOSTILE WITNESS (legal thriller, thriller) (The Witness Series,#1)	Rebecca Forster		y	6	y
B0031YJFCQ	The Girl Who Kicked the Hornet's Nest: Book 3 of the Millennium Trilogy (Vintage Crime/Black Lizard)	Stieg Larsson	Vintage	n	4	y
B0032JTTI0	A Cowboy For Christmas	Kristen James	Brilliant Book Press	n	6	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B003980ELA	The Skull Ring (Julia Stone)	Scott Nicholson	Haunted Computer Books	y	6	y
B003E74A66	The Prodigal Daughter	Jeffrey ARCHER	St. Martin's Paperbacks	n	6	y
B003F77EP4	Drummer Boy: A Supernatural Thriller (Sheriff Littlefield Series)	Scott Nicholson	Haunted Computer Books	y	6	y
B003I1WY2A	Water for Elephants: A Novel	Sara Gruen	Algonquin Books	n	4	y
B003K16PXC	Extremely Loud and Incredibly Close: A Novel	Jonathan Safran Foer	Houghton Mifflin Harcourt	n	3	y
B003NX7018	Jude Outlaw (Outlaw Lovers, Book One)	Jan Springer	Ellora's Cave	n	4	y
B003O86FMW	Catching Fire (The Second Book of the Hunger Games)	Suzanne Collins	Scholastic Press	n	4	y
B003RCJUCM	While the Savage Sleeps	Andrew E. Kaufman	Straightline Press	y	6	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B003SE7J6I	Beneath a Buried House: A Detective Elliot Mystery (Detective Elliot Mysteries)	Bob Avey	Deadly Niche Press	n	6	y
B003SNJZ02	Eagle's Run (Texas Passions, Book One)	Desiree Holt	Ellora's Cave	n	6	y
B003TU2JJ8	Captured	Victoria Lynne	Amazon.com	y	3	y
B003WJQ74E	Flowers for Algernon	Daniel Keyes	Houghton Mifflin Harcourt	n	4	y
B003WUYPPG	Unbroken: A World War II Story of Survival, Resilience, and Redemption	Laura Hillenbrand	Random House	n	5	y
B003XF1XOQ	Mockingjay (The Final Book of The Hunger Games)	Suzanne Collins	Scholastic Press	n	4	y
B003YFJ658	In Their Blood: A Novel	Sharon Potts	Oceanview Publishing	n	5	y
B003YFJ6G2	Frame-Up	John F. Dobbyn	Oceanview Publishing	n	4	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B003YL4LYI	A Dance with Dragons: A Song of Ice and Fire: Book Five	George R.R. Martin	Bantam	n	6	y
B003ZUYT3G	The Gates of Hell	Susan Sizemore	Speculation Press	n	4	y
B00408AOB8	The Baby Thief (A Romantic Thriller)	L.J. Sellers	Spellbinder Press	y	6	y
B0046A9V7I	SECRETS FROM THE DUST	George Hamilton	Browsing Rhino	y	5	y
B0047T7440	A Cup of Comfort Women of the Bible Devotional: Daily Reflections Inspired by Scripture's Most Beloved Heroines	James Stuart Bell	Adams Media	n	1	y
B00492CK1M	Spying in High Heels (High Heels Mysteries)	Gemma Halliday		y	5	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B004A90BXS	Heaven is for Real: A Little Boy's Astounding Story of His Trip to Heaven and Back	Todd Burpo	Thomas Nelson	n	5	y
B004BSGJFW	I'll Follow the Moon (Mom's Choice Award Honoree and Chocolate Lily Award Winner)	Stephanie Lisa Tara	Wee Words LLC	y	6	y
B004CFA9RS	Divergent	Veronica Roth	Katherine Tegen Books	n	6	y
B004CLYGGC	SILICONE	Carlos Meza	Roatan Press	y	5	y
B004DEPELY	The Paris Wife: A Novel	Paula Mclain	Ballantine Books	n	4	y
B004DERGVU	Falling for Rain	Janice Kirk		y	4	y
B004EHZXVQ	A Tale of Two Cities	Charles Dickens	Public Domain Books	n		n
B004HO5XJ8	Mr Right for the Night (Irish Romantic Comedy)	Marisa Mackle		y	3	n

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B004JN1D2I	George R. R. Martin's A Game of Thrones 4-Book Bundle	George R.R. Martin	Bantam	n	6	y
B004KKZ3GC	A Walk in the Snark	Rachel Thompson		y	5	y
B004LROUNG	The Litigators	John Grisham	Doubleday	n	4	y
B004NSVIPQ	Flight or Fancy?	Jude Ryan	Chelsea Square	y	2	y
B004QWZDKA	Crimson Leaf	SM Jonas		y	3	y
B004QX07EQ	Guilty Wives	James Patterson	Little, Brown and Company	n	5	y
B004QZ9VSM	Cake Mixes Cookbook (Gooseberry Patch)	Gooseberry Patch	Gooseberry Patch	y	6	y
B004SHNUGC	Rain	Leigh K. Cunningham	Vivante Publishing	y	6	y
B004SOIMOK	Live Organic (52 Brilliant Ideas)	Lynn Huggins-Cooper	Infinite Ideas (Trade)	y	4	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B004SY5QK8	Tracking Perception (a Becky McAllen Mystery Novel)	Cherri Galbiati	Cherri Galbiati	y	4	y
B004TGT3GI	The Devil Rogue	Lori Villarreal	Freeland Publishing	y	5	y
B004TQ8GP2	The Pursuit of God	A. W. (Aiden Wilson) Tozer		n		n
B004ULORYU	Killing Lincoln: The Shocking Assassination that Changed America Forever	Bill O'Reilly	Henry Holt and Co.	n	5	y
B004V3WT6K	Kill Shot (Mitch Rapp)	Vince Flynn	Atria/Emily Bestler Books	n	5	y
B004W2UBYW	Steve Jobs	Walter Isaacson	Simon & Schuster	n	5	y
B004XJRQUQ	The Hunger Games Trilogy	Suzanne Collins	Scholastic Press	n	4	y
B004YWK8WU	Convictions	Julie Morrigan		y	5	y
B004ZG8M3C	Return to Paradise	Carol Grace		y	5	y
B0050DIWFC	Defending Jacob: A Novel	William Landay	Delacorte Press	n	6	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B0050KTQ0K	Divorced, Desperate and Delicious (Divorced and Desperate)	Christie Craig	Love Spell	n	5	y
B0050ZKV8Q	The Highlander's Time	Belladonna Bordeaux	Decadent Publishing Company	n	4	y
B00514OWKY	Whispers in the Sand	Barbara Erskine	Sourcebooks Landmark	n	6	y
B0051CC7LC	Growing Up Amish	Ira Wagler	Tyndale House Publishers, Inc.	n	4	y
B0052RDH6K	The Next Always: Book One of the Inn BoonsBoro Trilogy	Nora Roberts	Berkley	n	4	y
B0052U9LXA	The Somali Doctrine	James Grenton	Intelligent Thriller Publishing	y	4	y
B0052VUNHC	Beautiful Disaster	Jamie McGuire	Jamie McGuire, LLC	y	5	n
B0053TIB6I	She Can Run	Melinda Leigh	Montlake Romance	n	5	y
B0053YNUCI	The Jesus Revolution: Learning from Christ's First Followers	Leith Anderson	Abingdon Press	n	2	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B0054QAY8I	Erasing Hell: What God Said about Eternity, and the Things We've Made Up	Francis Chan	David C. Cook	n	4	y
B0055PGUYU	The Power of Habit: Why We Do What We Do in Life and Business	Charles Duhigg	Random House	n	5	y
B00564GOM8	Victims: An Alex Delaware Novel	Jonathan Kellerman	Ballantine Books	n	5	y
B0056DRB0S	The Spy Who Left Me: An Agent Ex Novel (Agent Ex 1)	Gina Robinson	St. Martin's Paperbacks	n	6	y
B005723JSQ	I've Got Your Number: A Novel	Sophie Kinsella	The Dial Press	n	4	y
B0058CWYHC	Loving (Bailey Flanigan Series)	Karen Kingsbury	Zondervan	n	5	y
B0058UXHHK	Twenty-Eight and a Half Wishes (A Rose Gardner	Denise Grover Swank	Bramagioia Enterprises	y	6	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
	Mystery)					
B005BU9JK6	Gateway to Heaven	Beth Kery	Beth Kery Smashwords Edition	y	6	y
B005BUG6T8	Drift: The Unmooring of American Military Power	Rachel Maddow	Crown	n	5	y
B005BUG6TI	Betrayal: A Novel	Danielle Steel	Delacorte Press	n	6	y
B005C1N24I	1929 (Book One: Jonathan's Cross)	M.L. Gardner		y	6	y
B005E8287G	Healthy Breakfast Recipes	Alissa Carter		y	2	y
B005EJ3OPA	The Between Years	Derek Clendening		y	4	y
B005FFTRO0	Elephant Girl: A Human Story	Jane Devin		y	5	y
B005FLODDE	The Woman Who Wasn't There	Robin Gaby Fisher	Touchstone	n	5	y
B005GG0M0G	Mrs. Kennedy and Me	Lisa McCubbin	Gallery Books	n	5	y
B005GSYXYK	Stay Close	Harlan Coben	DUTTON ADULT	n	5	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B005GSZZ1A	Escape from Camp 14: One Man's Remarkable Odyssey from North Korea to Freedom in the West	Blaine Harden	VIKING ADULT	n	5	y
B005GSZZ1U	About That Night	Julie James	Berkley	n	6	y
B005GT0LZE	The Ghosts of Varner Creek	Michael Weems		y	2	y
B005HFHYM0	Private Games	James Patterson	Little, Brown and Company	n	5	y
B005IGVS6Q	Unfinished Business (Silhouette Intimate Moments)	Nora Roberts	Silhouette Special Releases	n	3	y
B005IGYXFE	The Lost Years	Mary Higgins Clark	Simon & Schuster	n	3	y
B005IQZB14	Wild: From Lost to Found on the Pacific Crest Trail	Cheryl Strayed	Knopf	n	5	y
B005J61DDI	The Righteous (Righteous Series #1)	Michael Wallace	Thomas & Mercer	n	4	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B005JMJ046	Love is Darkness (A Valerie Dearborn Novel)	Caroline Hanson		y	5	y
B005JPEG9M	Agartha's Castaway (First book in the Trapped In The Hollow Earth Saga)	Chrissy Peebles		y	5	y
B005JSV0ZW	Lone Wolf	Jodi Picoult	Atria/Emily Bestler Books	n	5	y
B005JWCCFU	Disappear, Love	E. Hughes	Love-Love Publishing	y	4	y
B005K0HDGE	11/22/63	Stephen King	Scribner	n	4	y
B005KWQ1U6	Three Moons Over Sedona	Sherry Hartzler	Rock House Publishing	y	5	y
B005L9B5YU	On the Island	Tracey Garvis-Graves		y	5	n
B005M662MM	FIRST DROP: Charlie Fox book four	Zoe Sharp	Murderati Ink	y	5	y
B005ML0EZS	Into the Black: Odyssey One [Remastered Edition]	Evan Currie	47North	n	5	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B005MT8PUQ	Forgotten God: Reversing Our Tragic Neglect of the Holy Spirit	Francis Chan	David C. Cook	n	5	y
B005MTBJ9A	Crazy Love: Overwhelmed by a Relentless God	Francis Chan	David C. Cook	n	4	y
B005MZN1HC	Imagine: How Creativity Works	Jonah Lehrer	Houghton Mifflin Harcourt	n	6	y
B005NKGEP2	The Expats: A Novel	Chris Pavone	Crown	n	4	y
B005O1BXOM	The Beginner's Goodbye	Anne Tyler	Knopf	n	3	y
B005OGJJ34	Breaking The Rules	Barbara Samuel		y	4	y
B005OUUTBG	Happy Birthday to Me Again (Birthday Trilogy, Book 2)	Brian Rowe	CreateSpace	y	5	y
B005PII27U	When the Walls Fell (Out of Time)	Monique Martin		y	5	y
B005S19K1G	Ties That Bind	Heather Huffman	Booktrope Editions	n	5	y
B005S8O9ZG	The Lifeboat:	Charlotte Rogan	Reagan Arthur Books	n	5	y

Amazon ASIN	Title	Author	Publisher	Self Published ?	Social Media Use Count <sup>25</sup>	Used in Regression?
	A Novel					
B005SJ1ATS	Jesus: Why the World Is Still Fascinated by Him	Tim LaHaye	David C. Cook	n	4	y
B005SZ0W14	Night Swim	Jessica Keener	The Fiction Studio	y	4	y
B005T634QC	Cloak (YA Fantasy)	James Gough	WiDo Publishing	n	3	y
B005TA7SFQ	Of Moths and Butterflies	V.R. Christensen	Captive Press	y	6	y
B005TNM736	Sleeping with Paris	Juliette Sobanet		y	5	y
B005UD1EJS	Capitol Murder	Phillip Margolin	Harper	n	2	y
B005UZGH7U	Guilt Trip (Blanco County Mysteries)	Ben Rehder		y	5	y
B005WUVLAG	Deadly Stillwater - Thriller (McRyan Mystery Series #2)	Roger Stelljes	Roger Stelljes	y	5	y
B005WZL0RK	Desperate Housedogs	Sparkle Abbey	Bell Bridge Books	n	5	y
B005Z57E18	The Rise and Fall of the Third Reich	William Shirer	RosettaBooks	n		n
B005ZJX468	Jump Cut	Lise McClendon	Thalia Press	y	3	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B0060NXUDK	Fact. Fact. Bullsh*t!: Learn the Truth and Spot the Lie on Everything from Tequila-Made Diamonds to Tetris's Soviet Roots - Plus Tons of Other Totally Random Facts from Science, History and Beyond!	Neil Patrick Stewart	Adams Media	n	3	y
B0061TCRKU	The Day No One Played Together: A Story About Compromise	Donalisa Helsley	Mirror Publishing	n	6	y
B0063KUGYG	The Westies: Inside New York's Irish Mob	T. J. English	MysteriousPress.com/Open Road	n	4	y
B006466CCE	Sendero	Max Tomlinson		y	3	y
B0067AN0VI	Calico Joe	John Grisham	Doubleday	n	4	y

Amazon ASIN	Title	Author	Publisher	Self Published ?	Social Media Use Count <sup>25</sup>	Used in Regression?
B006HWXKD4	The Lion, the Lamb, the Hunted: A Psychological Thriller	Andrew E. Kaufman	Straightline Press	y	6	y
B006ICVOUO	The Shoemaker's Wife	Adriana Trigiani	Harper	n	5	y
B006JTTGLU	Charming the Shrew (The Legacy of MacLeod)	Laurin Wittig	Montlake	n	6	y
B006JTTGSI	Daring the Highlander (The Legacy of MacLeod)	Laurin Wittig	Montlake	n	6	y
B006JTTJ08	In Search of Lucy: A Novel	Lia Fairchild	AmazonEncore	n	6	y
B006KHWEKQ	FEARLESS (King Series)	Tawdra Kandle	Amazon	y	5	y
B006KL37XU	The Price	Alexandra Sokoloff	Murderati Ink	y	6	y
B006KRYMM4	When Darkness Falls	R.G. Porter		y	6	y
B006KXITWW	Shelly's Second Chance (The Wish Granters, Book One)	L B Swan		y	1	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B006LU0GYS	Love Reborn: A Novel of the Black Dagger Brotherhood	J.R. Ward	NAL	n	5	y
B006MFT788	REVENGE IS SWEET	Patrice Wilton	Dreamscape Press	y	4	y
B006NFB5BY	Solo	Sarah Schofield	M.O.I. Publishing	y	6	y
B006OBU9DM	Wicked Werewolf Night - (Alluring Tales: Night Moves) (Werewolf Society)	Lisa Renee Jones		y	5	y
B006PNXK2G	Shattered	Karl S Jones		y	4	y
B006POB270	Elsie - Adventures of an Arizona Schoolteacher 1913-1916	Barbara Anne Waite	Palomar Mountain Bookworks	y	6	y
B006Q7N4PO	The Walking Man	Wright Forbucks		y	5	y
B006QRBIKM	No Dress Required	Cari Quinn	Entangled Publishing	n	6	y
B006RNBI6O	Thread of Hope	Jeff Shelby		y	5	y
B006TI7RY Y	Stuck With You	Trish Jensen	Bell Bridge Books	n	3	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B006V3E2PE	The Big Miss: My Years Coaching Tiger Woods	Hank Haney	Crown Archetype	n	4	y
B007250EN4	The Red Book	Deborah Copaken Kogan	Hyperion	n	5	y
B0077T2BHM	The Vow: The True Events that Inspired the Movie	Krickitt Carpenter	B&H Books	n	2	y
B00790TI0W	The Marriage Bargain (Marriage to a Billionaire)	Jennifer Probst	Entangled Publishing, LLC	n	5	y
B00794UQ9K	No One But You	Jillian Hart	Harlequin Special Releases	n	3	y
B007A1V23Q	A Texan's Honor (Heart of a Hero)	Shelley Gray	Abingdon Press Fiction	n	4	y
B007A4V33M	An Unexpected Twist (Kindle Single)	Andy Borowitz		n	5	y
B007C81VNI	Three Days in Seattle, a Novel of Romance and Suspense	Debra Burroughs	Lake House Books	y	5	y
B007C8EYXC	VIOLENCE	Timothy McDougall		y	4	y

<b>Amazon ASIN</b>	<b>Title</b>	<b>Author</b>	<b>Publisher</b>	<b>Self Published ?</b>	<b>Social Media Use Count<sup>25</sup></b>	<b>Used in Regression?</b>
B007DH9C3E	Healing Hearts	Kim Watters		y	5	y
B007ECU5TI	Grow your own food	Infinite Ideas	Infinite Ideas (Trade)	y	1	y
B007EFHQHE	The Mother Road	Jennifer ALee	Abingdon Press	n	6	y
B007GR28XC	FERAL SINS (The Phoenix Pack)	Suzanne Wright	Suzanne Wright	y	3	y
B007HH5IO2	Blind Courage	Bill Irwin	Bill Irwin	y	2	y
B007IXWKUK	Fifty Shades Darker: Book Two of the Fifty Shades Trilogy	E L James	Vintage	n	6	y
B007IXWL2C	Fifty Shades Freed: Book Three of the Fifty Shades Trilogy	E L James	Vintage	n	6	y
B007J4T2G8	Fifty Shades of Grey: Book One of the Fifty Shades Trilogy	E L James	Vintage	n	6	y
B007JC2942	Playbook 2012: Inside the Circus--Romney, Santorum and	Evan Thomas	Random House	n	1	y

Amazon ASIN	Title	Author	Publisher	Self Published ?	Social Media Use Count <sup>25</sup>	Used in Regression?
	the GOP Race (POLITICO Inside Election 2012)					
B007JMIENC	The Girl Who Came Home - A Titanic Novel	Hazel Gaynor		y	5	y
B007JW1UHY	Vampire Dawn (Vampire for Hire #5)	J.R. Rain		y	5	y
B007LBD71A	Lifeboat No. 8: An Untold Tale of Love, Loss, and Surviving the Titanic (Kindle Single)	Elizabeth Kaye	Byliner Inc.	y	5	y
B007MQZ9J2	Basic Training (Kindle Single)	Kurt Vonnegut	RosettaBooks	n	3	y
B007MTR4AQ	Ball Four (RosettaBooks Sports Classics)	Jim Bouton	RosettaBooks	n	4	y
B007O0AYUA	CLUB JUSTICE (The Trinity Falls Series)	Mara McBain		y	6	y

Amazon ASIN	Title	Author	Publisher	Self Published ?	Social Media Use Count <sup>25</sup>	Used in Regression?
B007QEE6YY	City of Darkness (City of Mystery)	Kim Wright		y	5	y
B007QVABM8	The Hunger Games #2: Catching Fire (Discussion Guide)	Suzanne Collins	Scholastic	n	4	y

**Error! Not a valid link.**

## WORKS CITED

- (2011). Looking for the 50% solution. *Publishers Weekly*, (Dec. 30, 2011)
- Abdesslem, F. B., Parris, I., & Henderson, T. (2012). Reliable online social network data collection. In A. Abraham (Ed.), *Computational Social Networks. Volume 3, Mining and Visualization* doi:<http://dx.doi.org/10.1007/978-1-4471-4054-2>
- Allen, T. J. (1977). *Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization*. Cambridge, Mass: MIT Press.
- Americaneditor. (2010). The eBook wars: The gatekeeper role. Retrieved January 20, 2012, from <http://americaneditor.wordpress.com/2010/01/19/the-ebook-wars-the-gatekeeper-role/>
- Americaneditor. (2011). Gatekeeping: Necessary or not in the eBook era. Retrieved January 20, 2012, from <http://americaneditor.wordpress.com/2011/04/18/gatekeeping-necessary-or-not-in-the-ebook-era/>
- Bagrow, J. P., & Ben-Avraham, D. (2005). On the google-fame of scientists and other populations. *Modeling Cooperative Behavior in the Social Sciences*, 779 81-89. doi:10.1063/1.2008594

- Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web*, Lyon, France. 519-528. doi:10.1145/2187836.2187907
- Barzilai-Nahon, K. (2008). Toward a theory of network gatekeeping: A framework for exploring information control. *Journal of the American Society for Information Science and Technology*, 59(9), 1493-1512. doi:10.1002/asi.20857
- Barzilai-Nahon, K. (2009). Gatekeeping: A critical review. *Annual Review of Information Science and Technology*, 43(1), 1-79. doi:10.1002/aris.2009.1440430117
- Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5), 407-424.
- Beheshti, J. (1992). Browsing through public access catalogs. *Information Technology & Libraries*, 11(3), 220-228.
- Bennet, L. (2006). *Ebooks: The options: A manual for publishers*. London, England: The Publishers Association.
- Bennett, L. (2011). Ten years of e-books: A review. *Learned Publishing*, 24(3), 222-229. doi:10.1087/20110310
- Biba, P. (2011). Quality is the new gatekeeper: How ebooks have changed my reading. Retrieved January 20, 2012, 2012, from <http://www.teleread.com/paul-biba/quality-is-the-new-gatekeeper-how-ebooks-have-changed-my-reading/>

Blake, V. L. (1989). The role of reviews and reviewing media in the selection process.

*Collection Management*, 11(1), 1-40.

Bodoff, D. (2006). Relevance for browsing, relevance for searching. *Journal of the*

*American Society for Information Science and Technology*, 57(1), 69-86.

Boog, J. (2011). Top 20 facebook apps for book lovers. Retrieved December 15, 2012,

from [http://www.mediabistro.com/galleycat/top-20-book-focused-facebook-apps\\_b28441](http://www.mediabistro.com/galleycat/top-20-book-focused-facebook-apps_b28441)

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the

social sciences. *Science*, 323(5916), 892-895. doi:10.1126/science.1165821

Bosman, J. (2012, 3/9/2012). Discreetly digital, erotic novel sets American women

abuzz. *The New York Times*

Bowker (2012a). New book titles and editions, 2002-2011. Retrieved September 15,

2012, from

[http://www.bowker.com/assets/downloads/products/isbn\\_output\\_2002-2011.pdf](http://www.bowker.com/assets/downloads/products/isbn_output_2002-2011.pdf)

Bowker (2012b). Self-publishing sees triple-digit growth in just five years, says Bowker.

Retrieved May 10, 2013, from [http://www.bowker.com/en-](http://www.bowker.com/en-US/aboutus/press_room/2012/pr_10242012.shtml)

[US/aboutus/press\\_room/2012/pr\\_10242012.shtml](http://www.bowker.com/en-US/aboutus/press_room/2012/pr_10242012.shtml)

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal*

*Statistical Society. Series B (Methodological)*, 26(2), pp. 211-252.

- Bradley, J., Fulton, B., & Helm, M. (2012). Self-published books: An empirical "snapshot". *Library Quarterly*, 82(2), 107-140.
- Bradley, J., Fulton, B., Helm, M., & Pittner, K. (2011). Non-traditional book publishing. *First Monday*, 16(8)
- Breedt, A., & Walter, D. (2012). *Whitepaper: The link between metadata and sales*. Surrey, England: Nielsen Bookscan.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117.  
doi:10.1016/S0169-7552(98)00110-X
- Bry, I., & Afferbach, L. (1961). Book reviewing in the sciences of human behavior as a contribution to scholarship by the scientific community. *Mental Health Book Review Digest*, 6(July), i-viii.
- Brynjolfsson, E., Smith, M., & Hu, Y. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 49(11), 1580-1596.
- Bui, A. (2012). The challenges of discovering online research / reference content: An introduction to the end user's perspective. In S. Polanka (Ed.), *E-reference context and discoverability in libraries: Issues and concepts* (pp. 19-33). Hershey, PA: IGI Global.

- Bui, C. (2010). How online gatekeepers guard our view -- news portals' inclusion and ranking of media and events. *Global Media Journal: American Edition*, 9(16), 1-41.
- Bush, V. (1945). As we may think. *The Atlantic*, July, 1945
- Butts, C. T. (2008). Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, 11, 13-41. doi:/j.1467-839X.2007.00241.x
- Chandler, O. (2012). How people discover books online. Retrieved January 15, 2013, from <http://www.slideshare.net/PatrickBR/goodreads-how-people-discover-books>
- Charman-Anderson, S. (2013). Can Nielsen BookScan stay relevant in the digital age? *Forbes*, January 7, 2013
- Chevalier, J., & Goolsbee, A. (2003). Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. *Quantitative Marketing and Economics*, 1, 203-222.
- Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), 345-354.
- Coleman, J. (2007). Browsing 101: How do you find a good book? *Library Media Connection*, 25(4), 42-43.
- Coser, L. A. (1975). Publishers as gatekeepers of ideas. *The Annals of the American Academy of Political and Social Science*, 421(1), 14-22.  
doi:10.1177/000271627542100103

- Crane, D. (1967). The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 2(4), 195-201.
- Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(10), 1577-1593.
- Denton, W. (2007). FRBR and the history of cataloging. In A. G. Taylor (Ed.), *Understanding FRBR: What it is and how it will affect our retrieval*. Westport, CT: Libraries Unlimited. doi:<http://hdl.handle.net/10315/1250>
- Deutsch, P. (1991). Resource discovery in an internet environment - the archie approach. *Information and Knowledge Management*, 2(1), 45-51. doi:10.1108/eb047253
- Dhanasobhon, S., Chen, P., Smith, M. D., & Chen, P. (2007). An analysis of the differential impact of reviews and reviewers at amazon.com. *ICIS*, Montreal, Quebec, Canada. 94.
- Digital Book Wire. (2012). Print book and ebook sales feed into each other. Retrieved January 15, 2013, from <http://www.digitalbookworld.com/2012/print-book-and-e-book-sales-feed-into-each-other-says-lulu-com/>
- Dilevko, J. & Dali, K. (2006). The self-publishing phenomenon and libraries. *Library & Information Science Research*, 28(2), 208-234.
- Ericson, J. (2009). Metator, librarian, gatekeeper, broker. Retrieved January 20, 2012, from [http://www.information-management.com/blogs/metadata\\_editor\\_metador-10016283-1.html](http://www.information-management.com/blogs/metadata_editor_metador-10016283-1.html)

- Esposito, J. (2011). The vexed problem of libraries, publishers and ebooks. Retrieved January 20, 2012, from <http://scholarlykitchen.sspnet.org/2011/03/21/the-vexed-problem-of-libraries-publishers-and-e-books/>
- Fenner, T., Levene, M., & Loizou, G. (2010). Predicting the long tail of book sales: Unearthing the power-law exponent. *Physica A: Statistical Mechanics and its Applications*, 389(12), 2416-2421. doi:10.1016/j.physa.2010.02.021
- Flamm, M. (2009). 367 magazines shuttered in 2009. Retrieved January 15, 2010, from <http://www.crainsnewyork.com/article/20091211/FREE/912119988>
- Flood, A. (2011). Hardback sales plummeting in age of the ebook. *The Guardian* August 12, 2011
- Foster, A., & Ford, N. (2003). Serendipity and information seeking: An empirical study. *Journal of Documentation*, 59(3), 321-340.
- Fulton, B. (2010a). *Enhancing browse with social metadata*. Unpublished manuscript.
- Fulton, B. (2010b). *Book reviews and collection development: A crisis of numbers*. Unpublished manuscript.
- Fulton, B. (2010c). *Online reputation and consumer reviews: The influence of impression management tactics and hyperpersonal effects*. Unpublished manuscript.
- Fulton, B. (2009d). *Gaming the system: Manipulation of consumer reviews*. Unpublished manuscript.

Fulton, B. (2009e). *Why scholars read scholarly book reviews: A uses and gratifications approach*. Unpublished manuscript.

Geiber, W. (1956). Across the desk: A study of 16 telegraph editors. *Journalism Quarterly*, 33(4), 423-432.

Gladwell, M. (2000). *The tipping point: How little things can make a big difference* (1st ed.). Boston: Little, Brown.

Glogoff, S. (1988). Reviewing the gatekeepers: A survey of referees of library journals. *Journal of the American Society for Information Science*, 39(6)

Grabois, A. (2007). The decline and demise of the newspaper book review. Retrieved October 4/2009, from <http://www.beneaththecover.com/2007/11/19/the-decline-and-demise-of-the-newspaper-book-review/>

Granovetter, M. S. (1973). The strength of weak ties. *The American Journal of Sociology*, 78(6), 1360-1380.

Gravano, L., Garcia-Molina, H., & Tomasic, A. (1993). *The efficacy of GLOSS for the text database discovery problem*. Stanford InfoLab; Stanford University.

Greco, A. N. (2005). *The book publishing industry* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates.

Greco, A. N. (2011). *The book publishing industry*. London: Routledge

Greco, A. N., Rodriguez, C. E., & Wharton, R. M. (2007). *The culture and commerce of publishing in the 21st century*. Stanford, CA: Stanford Business Books.

- Greenfield, J. (2012). ALA appeals to publishers to sell ebooks to libraries. Retrieved March 14, 2012, 2012, from <http://www.digitalbookworld.com/2012/head-of-library-association-appealsto-publishers-to-sell-e-books-to-libraries/>
- Griffiths, Jillian and Brophy, Peter. (2005). Student searching behavior and the web: Use of academic resources and google. *Library Trends*, 53(4), 539-554.
- Harmon, A. (2004). Amazon glitch unmasks war of reviewers. Retrieved December 5, 2010, from <http://www.nytimes.com/2004/02/14/technology/14AMAZ.html>
- Horrigan, J. (2008a). *Online shopping*. Washington, D.C.: Pew Internet & American Life Project.
- Horrigan, J. B. (2008b). *The internet and consumer choice: Online Americans use different search and purchase strategies for different goods*. Pew Internet and American Life Project.
- Hoskins, A., & O'Loughlin, B. (2011). Remediating jihad for western news audiences: The renewal of gatekeeping? *Journalism*, 12(2), 199-216.  
doi:10.1177/1464884910388592
- Hu, N., Liu Jr., L., Sambamurthy, V., & Chen, B. (2009). Are online reviews just noise? The truth, the whole truth, or only the partial truth? *Research Collection School of Information Systems (Open Access), Paper 338* Retrieved December 15, 2012 from [http://ink.library.smu.edu.sg/sis\\_research/338](http://ink.library.smu.edu.sg/sis_research/338)
- Hu, N., Liu Jr., L., & Zhang, J. (2008). *Do online reviews affect product sales? The role of reviewer characteristics and temporal effects*. Unpublished manuscript.

- Hu, N., Liu, L., Bose, I., & Shen, J. (2010). Does sampling influence customers in online retailing of digital music? *Information Systems and eBusiness Management*, 8(4), 357-377. doi:10.1007/s10257-009-0116-6
- Iba, T., Yoshida, M., Fukami, Y., & Saitoh, M. (2008). Power-law distribution in Japan's book sales market. Japan.
- Ingram, H. M., & Mills, P. B. (1989). Reviewing the book reviews. *PS: Political Science and Politics*, 22(3), 627-636.
- Interpreting regression coefficients. (2013). Retrieved June 15, 2013, from <http://www.theanalysisfactor.com/interpreting-regression-coefficients/>
- Janetzko, D. (2008). Objectivity, reliability and validity of search engine count estimates. *International Journal of Internet Science*, 3(1), 7-33.
- Joyce, D. F. (1998). Unique gatekeepers of black culture: Three black librarians as book publishers. *Untold Stories: Civil Rights, Libraries and Black Librarianship*. Edited by John Mark Tucker. Urbana-Champaign, Illinois: Illinois University at Urbana Champaign, Graduate School of Library and Information Science, 1998, p.151-5
- Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *The Public Opinion Quarterly*, 21, 61-78.
- Katz, E., Lazarsfeld, P. F., & Columbia University. Bureau of Applied Social Research. (1955). *Personal influence; the part played by people in the flow of mass communications*. Glencoe, Ill.: Free Press.

- Kaufman, L. (2013). Bookish, new web site, provides information on books and authors. *New York Times*. February 4, 2013
- Kay, A. (1972). A personal computer for children of all ages. *ACM National Conference*, Boston, MA. , *Boston*
- Keen, A. (2007). *The cult of the amateur: How today's internet is killing our culture* (1st ed.). New York: Doubleday/Currency.
- Kilborn, P. (2010). Identification of e-books. *Learned Publishing*, 23(2), 166-168.  
doi:10.1087/20100215
- Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, 33(1), 147-151. doi:10.1162/coli.2007.33.1.147
- Kurtz, H. (2009) Post to end stand-alone book section. Retrieved November 27, 2009, from <http://www.washingtonpost.com/wp-dyn/content/article/2009/01/28/AR2009012802208.html>
- Lang, B. (2012). Movie buzz more likely to spread by word-of-mouth than social networks. Retrieved October 2, 2012, from <http://movies.yahoo.com/news/movie-buzz-more-likely-spread-word-mouth-social-223949458.html>
- Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The people's choice: How the voter makes up his mind in a presidential campaign*. New York: Duell, Sloan, and Pearce.
- Lebert, M. (2009). *A short history of e-books*. Chapel Hill, NC: Project Gutenberg.

- Ledbetter, J. (2010). Kindlerotica: The strange but inevitable rise of e-reader pornography. Retrieved December 15, 2012, from <http://www.slate.com/articles/technology/technology/2010/09/kindlerotica.html>
- Levine-Clark, M., & Jobe, M. M. (2007). Do reviews matter? An analysis of usage and holdings of choice-reviewed titles within a consortium. *The Journal of Academic Librarianship*, 33(6), 639-646. DOI: 10.1016/j.acalib.2007.09.002
- Lewin, K. (1943). Forces behind food habits and methods of change. *Bulletin of the National Research Council*, 108, 35-65.
- Lewin, K. (1947). Frontiers in group dynamics. *Human Relations*, 1(2), 143-153.  
doi:10.1177/001872674700100201
- Lichtenberg, J. (2011). In from the edge: The progressive evolution of publishing in the age of digital abundance. *Publishing Research Quarterly*, 27(2), 101-112.  
doi:10.1007/s12109-011-9212-9
- Lim, M. (1995). Gatekeepers: What are they? *Link-Up*, March 1995, 18-20.
- Lin, T. M., Huang, Y. K., & Yang, W. I. (2007). An experimental design approach to investigating the relationship between internet book reviews and purchase intention. *Library and Information Science Research*, 29, 397-415.
- Lin, T. M. Y., Luarn, P., & Huang, Y. K. (2005). Effect of internet book reviews on purchase intention: A focus group study. *The Journal of Academic Librarianship*, 31(5), 461-468. doi:DOI: 10.1016/j.acalib.2005.05.008

- Lucey, B. (2009). The New York Times: A chronology: 1851 - 2006. Retrieved November 20, 2009, from <http://www.nysl.nysed.gov/nysnp/nytlucey.htm>
- Manzi, J. (2012). *Uncontrolled: The surprising payoff of trial-and-error for business, politics, and society*. New York: Basic Books.
- Mayer, B. (2011). The real gatekeepers in publishing now? Authors. Retrieved January 20, 2013 from <http://writeitforward.wordpress.com/2011/09/14/the-real-gatekeepers-in-publishing-now-authors/>
- McCutcheon, R. P. (1922). The beginnings of book-reviewing in English periodicals. *PMLA*, 37(4), 691-706.
- McDonald, J. H. (2009). *Handbook of biological statistics: 2nd edition*. Baltimore, MD: Sparky House Publishing.
- Milliot, J., & Coffey, M. (2010). Self-publishing comes of age. *Publisher's Weekly*, 257(50), 1-2.
- Multiple linear regression model. (2011). Retrieved June 15, 2013, from <http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=0CD4QFjAC&url=http%3A%2F%2Fwritingcenter.waldenu.edu%2FDocuments%2FMedia-Center%2FMultipleLinearRegressionAnalysis.ppt>
- Nakov, P., & Hearst, M. (2005). A study of using search engine page hits as a proxy for n-gram frequencies. *Proceedings of Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

NIST/SEMATECH. (2012). e-handbook of statistical methods. Retrieved April 30, 2013,

from <http://www.itl.nist.gov/div898/handbook/eda/section3/boxcoxno.htm>

Okerson, A., & Mogge, D. (1994). In Okerson A., Mogge D. (Eds.), *Gateways, gatekeepers, and roles in the information omniverse: Proceedings of the third symposium: November 13-15, 1993, the Washington Vista Hotel, Washington, DC*. Washington, DC: Association of Research Libraries, Office of Scientific and Academic Pub.

Onnela, J., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Barabási, A.

(2007). Structure and tie strengths in mobile communication networks.

*Proceedings of the National Academy of Sciences*, 104(18), 7332-7336.

doi:10.1073/pnas.0610245104

Pham, A. (2010, 12/26/2010). Book publishers see their role as gatekeepers shrink. *Los*

*Angeles Time*

Pinto, C. M. A., Mendes Lopes, A., & Machado, J. A. T. (2012). A review of power laws

in real life phenomena. *Communications in Nonlinear Science and Numerical*

*Simulation*, 17(9), 3558-3578. doi:10.1016/j.cnsns.2012.01.013

Pollard, J. (2007). Google result counts are a meaningless metric. Retrieved December

15, 2012, from

<http://homepage.ntlworld.com/jonathan.deboynepollard/FGA/google-result-counts-are-a-meaningless-metric.html>

- Rainie, L., Zickuhr, K., Purcell, K., Madden, M., & Brenner, J. (2012). *The rise of e-reading*. Washington, D.C.: Pew Research Center.
- Reitz, J. M. (2004). *Dictionary for library and information science*. Westport, Conn: Libraries Unlimited.
- Robinson, C. (2010). The trouble with Amazon. *The Nation*, August 2/9, 2010
- Rogers, E. M. (1962). *Diffusion of innovations*. New York: Free Press of Glencoe.
- Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York: Free Press.
- Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.
- Rorvig, M. E. (1988). How do you browse? *Library Journal*, 113(1), 61.
- Rosenbaum, S. C. (2011). *Curation nation: How to win in a world where consumers are creators*. New York: McGraw-Hill.
- Rosenthal, M. (2012). Amazon sales ranking and author rank. Retrieved December 15, 2012, from <http://www.fonerbooks.com/surfing.htm>
- Rubinfeld, D. (2011). Reference guide on multiple regression. *Reference manual on scientific evidence: Third edition* (pp. 303-358). Washington, DC: The National Academies Press.
- Sakia, R. (1992). The box-cox transformation technique: A review. *The Statistician*, 41, 169-178.

- Sathe, N., Grady, J., & Guise, N. (2002). Print versus electronic journals: A preliminary investigation into the effect of journal format on research processes. *Journal of the Medical Library Association*, 90(2), 235-243.
- Schmitt, G. (2007). Razorfish digital consumer behavior study. Retrieved September 28, 2007, from <http://www.digitaldesignblog.com/2007/09/28/digital-consumer-behavior-study/>
- Serebnick, J., & Cullars, J. (1980). An analysis of publishers of books reviewed in key library journals. *Library and Information Science Research*, 6(3), 289-303.
- Serebnick, J. (1981). Book reviews and the selection of potentially controversial books in public libraries. *The Library Quarterly*, 51(4), 390-409.
- Shearer, K. D. (1981). The selector as gatekeeper. *Public Libraries*, 20, 91-93.
- Shearer, K. D. (1983). Applying new theories to library selection. *Drexel Library Quarterly*, 19(2)
- Shoemaker, P. J. (1991). *Gatekeeping*. Newbury Park: Sage Publications.
- Shoemaker, P. J., Eichholz, M., Kim, E., & Wrigley, B. (2001). Individual and routine forces in gatekeeping. *Journalism & Mass Communication Quarterly*, 78(2), 233-246. doi:10.1177/107769900107800202
- Shoemaker, P. J., & Vos, T. P. (2009). *Gatekeeping theory*. New York: Routledge.
- Skerik, S. (2011). *Unlocking social media for pr*. New York, NY: PR Newswire Association.

- Sorensen, A. T., & Rasmussen, S. (2004). *Is any publicity good publicity? A note on the impact of book reviews*. Unpublished manuscript.
- Streitfeld, D. (2011, 8/12/2011). Amazon cracks down on some e-book 'publishers'. *The New York Times*
- Streitfeld, D. (2012). Cut in ebook pricing by Amazon is set to shake rivals. *New York Times*, April 11, 2012
- Sturges, P. (2001). Gatekeepers and other intermediaries. *ASLIB Proceedings*, 53(2), 62-67.
- Su, C., & Contractor, N. (2011). A multidimensional network approach to studying team members' information seeking from human and digital knowledge sources in consulting firms. *Journal of the American Society for Information Science and Technology*, 62(7), 1257-1275. doi:10.1002/asi.21526
- Sullivan, H. A. (1958). Vanity press publishing. *Library Trends*, 7(1), 105-115.
- Thompson, J. B. (2005). *Books in the digital age: The transformation of academic and higher education publishing in Britain and the United States*. Cambridge, U.K.; Malden, Mass: Polity Press.
- Thompson, J. B. (2010). *Merchants of culture: The publishing business in the twenty-first century*. Cambridge, UK; Malden, MA: Polity.
- Tian, X., & Martin, B. (2011). Impacting forces on eBook business models development. *Publishing Research Quarterly*, 27(3), 230-246. doi:10.1007/s12109-011-9229-0

- Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32(4), 425-443.
- Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, 35(4), 469-480. doi:10.1177/0165551509103598
- Vassiliou, M., & Rowley, J. (2008). Progressing the definition of "e-book". *Library Hi Tech*, 26(3), 355-368. doi:10.1108/07378830810903292
- Verso Advertising. (2012). 2011 survey of book-buying behavior. Retrieved January 15, 2013, from <http://www.versoadvertising.com/DBWsurvey2012/>
- Walker, J. (2009). New resource discovery mechanisms. *The E-Resources Management Handbook*, , 78-89. doi:10.1629/9552448-0-3-8.1
- Wasserman, S. (2007). Goodbye to all that. *Columbia Journalism Review*, 46(3), 43-54.
- Weir, D. (2011). Ebook leader smashwords CEO: "we've eliminated the publisher as gatekeeper". Retrieved January 20,, 2012, from <http://www.7x7.com/tech-gadgets/ebook-leader-smashwords-ceo-weve-e>
- Westley, B. H., & MacLean, M. S. A conceptual model for communications research. *Journalism Quarterly*, 34, 31-38.
- Wharton, R. M., & Greco, A. N. (2004). Small and independent publishers: An analysis of sales data. Retrieved November 9, 2009, from <http://www.ibpa-online.org/articles/shownews.aspx?id=2017>

- White, D. M. (1950). The "gate keeper": A case study in the selection of news. *Journalism Quarterly*, 27, 383-390.
- White, D. M. (1964). The 'gatekeeper': A case study in the selection of news. *People, society, and mass communications*. (pp. 160-172). [New York: Free Press of Glencoe.
- Whitworth, B., & Friedman, R. (2009). Reinventing academic publishing online. part II: A socio-technical vision. *First Monday*, 14(9)
- Wyatt, E. (2004, 10/5/2004). An honest book review from Kirkus? only \$350. *The New York Times*
- Yu, H., & Young, M. (2004). The impact of web search engines on subject searching in OPAC. *Information Technology and Libraries*, 23(4), 168-180.
- Zhao, J., Wu, J., & Xu, K. (2010). Weak ties: Subtle role of information diffusion in online social networks. *Phys.Rev.E*, 82(1), 016105.  
doi:10.1103/PhysRevE.82.016105
- Ziarko, W. (1995). Introduction to the special issue on rough sets and knowledge discovery. *Computational Intelligence*, 11(2), 223-226. doi:10.1111/j.1467-8640.1995.tb00028.x