eBooks, Authors and Social Media Outreach:

Social Gatekeeping, Discovery and the Serendipitous Tie*

By

Bruce Fulton, Ph.D.

This document supplements the video presentation with copies of the slides and a full script of the text

Video Presentation Link: http://www.youtube.com/watch?v=kPqL279M-tE

Originally presented 7/19/2013 at the University of Arizona

This document and the accompanying video linked above are released under a creative commons license as described below. You may copy or distribute these works for non-commercial use in their original unaltered form.

Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License.

Contact the author at bfulton@email.arizona.edu

http://sirls.arizona.edu/bruce-fulton

* adapted from:

Contents

Presentation Script .............................................................................................................. 4

Slide 1 – Title Slide ......................................................................................................... 4

Slide 2 - Introduction ...................................................................................................... 5

Slide 3 – The Discovery Problem ................................................................................... 7

Slide 4 - Purposes ............................................................................................................ 9

Slide 5 – Foundations of Social Gatekeeping ............................................................... 10

Slide 6 – The Serendipitous Tie .................................................................................... 14

Slide 7 – Conventional Weak Tie Diffusion ................................................................. 15

Slide 8 - Serendipitous Tie Diffusion ........................................................................... 16

Slide 9 – Social Gatekeeping Defined .......................................................................... 18

Slide 10 – The Traditional Publishing Chain Re-envisioned ........................................ 19

Slide 11 – Today’s Publishing Chain ........................................................................... 21

Slide 12 – Social Gatekeeping as Strategy ................................................................... 22

Slide 13 – Nature of the Study ...................................................................................... 23

Slide 14 – Dependent Variables .................................................................................... 26

Slide 15 – Independent Variables ................................................................................ 27

Slide 16 - Methodology ................................................................................................ 28

Slide 17 – Phase I Highlights ........................................................................................ 30

Slide 18 – Selected Descriptive Statistics ..................................................................... 32

Slide 19 – No-Sales and Social Media Use .................................................................. 33

Slide 20 – Phase II and Regression ............................................................................... 34

Slide 21 – Phase II – Key Regression Statistics .............................................................. 36
Welcome, and thanks for your interest. Today I will summarize key findings of my dissertation research and present:

- the overarching purpose of the research I conducted,
- the theoretical framework I developed that extends existing theory and puts the research into context,
- the specific questions and hypotheses I developed that the research addresses,
- the methodology I used to conduct the research,
- the results of the research and interpretation of those results, and
- the importance of this work and its greater potential impact, and directions for future investigation.
The book publishing landscape changed dramatically during the first decade of the 21st century. These changes are driven by advances in technology, the evolution of computer mediated communication channels, and marketing forces that continue to concentrate mainstream publishing houses into mega-conglomerates while at the same time spawning thousands of niche market players and perhaps hundreds of thousands of self-publishers.

Today, people look not only to mass media for consumer information but also social media where recommendations come from friends and family or like-minded consumers. In the meantime, traditional sources of reviews and advertisements such as newspapers and magazines continue to decline.

Technology has made it easy for authors to self-publish, and new business models such as print on demand and computer-based authoring tools have eliminated barriers of printing, warehousing and sales outlets.

EBooks, many of which may never be printed, now account for 20% or more of sales for major publishers. There is a hyper-abundance of choice and a heterogeneous marketplace
comprised not only of bricks and mortar stores but online marketplaces with virtually unlimited choices.
In 2011, at least 1.5 million new titles were published in the United States alone, a 400% increase in just five years, and the number of self-published titles has eclipsed titles produced by mainstream publishers. The full count is in fact unknown and virtually unknowable since large numbers of new titles never acquire an ISBN number.

Only a portion of sales of books, and especially of eBooks, are tracked by companies such as Nielson, and then only sales from major outlets are counted. Only a small number of titles are reviewed annually by the trade and popular press – perhaps 25k to 30k at most – and opportunities for previously common sources of book browse such as displays in bricks and mortar stores are becoming less common.

How do readers find the books they want to read, given the number to choose from, and how do readers navigate the diffuse and disparate sources of information about them?

Information about books is moving online, and marketing for books is moving online as well.

According to many marketing surveys, friends and family are among the important sources for book recommendations, along with computer mediated social channels, advertising,
and various forms of search and automated recommendation systems. Many of these interactions occur today within online social networks.

Traditional saturation mass media marketing is out of the price range of all but the likely best sellers from major publishing conglomerates. Authors today – self-published and traditionally published alike – are proactively taking action in order to connect readers to their books, and even traditional publishers with media budgets expect authors to shoulder some of the responsibility for getting their books into the hands of readers.
Slide 4 - Purposes

Purposes

• Describe How Authors Use Social Media
• Determine if and to what extent this increases discoverability and readership
• Social Gatekeeping as a Framework of Understanding

The research I’m going to describe to you today looks at how authors use social media, and how authors use social media as a strategy to connect readers to books.

In particular, the research looks generally at samples of current literary output and the kinds of social media outreach the authors of those works are using to connect to potential readers, if any. The research provides a first look and estimation of the relative impact of those strategies on discovery and readership and lays a foundation of understanding for how these strategies might work.
The name I’ve given to the framework I’ve developed is Social Gatekeeping, and I’ll start by describing at least briefly its foundations in theories of communication, marketing, social network theory and dyadic communication, and library and information science. Social gatekeeping extends these theories to support my research.

1. I’ll first mention the traditional publishing chain as described by book industry expert John B. Thompson.¹ This view of publishing draws from the management and marketing literature on commercial supply and value chains.

The publishing chain consists of 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. So we start with the author who creates the content and follow it along to the publisher where various intermediaries add value, such as copy-editing, design, typesetting,

---

proof-reading, printing and binding, sales and marketing, warehousing and distribution and finally to book sellers who get the works into the hands of readers.

Through the end of the 20th century, this supply chain is the principle way books made their way from author to publisher. There were few alternatives. There was so-called Vanity Self-Publishing, but vanity publishing lacked access to the publishing chain and very few books published that way found readership. As traditionally conceived and illustrated by Thompson, the publishing supply chain relies principally on mass media, including both marketing and reviews, to initiate information flow and promote discovery and sales.

2. Gatekeeping was first proposed by Kurt Lewin in 1947\(^2\) and as initially presented envisioned channels through which food came to table, such as from the garden or from the market. Gatekeepers are the decision makers at key points in the channels, called information gates, where critical decisions about what might pass through the gate are made. Lewin found that the key to influencing what made it to the table was to influence the gatekeeper.

Gatekeeping has since been adopted by many disciplines as a theory, framework or model. In the field of journalism, gatekeeping by editors is posited as the mechanism by which millions of messages are filtered, modified and transformed into the few that make their way to readers. This is primarily an act of information filtered out.

Library science also recognizes gatekeeping in the literature but in a different context. Librarians and other information professionals intermediate information and curate it by finding the best work and making it available. In this sense, gatekeeping is a positive force for discovery and knowledge transfer, and this is primarily an act of information filtered in.

So 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.

3. Once information reaches someone within a social network, say through mass media advertising, social network theories and communication theories describe how information diffuses through those networks.

I’ll mention the two-step flow of communication theories, developed by Paul Lazarsfeld in 1944 and subsequently by Elihu Katz beginning in 1955. These theories were proposed as a counter to early theories about mass media that posited that mass media had direct and powerful effects on individuals. The two-step and multi-step flow theories suggested instead that mass media reached certain people who were then responsible for influencing others in their social circles.

I’ll also mention Everitt Rogers’ Diffusion of Innovations work first presented in 1962 with several editions since then. Rogers writes that the first two phases of innovation adoption are knowledge (which is discovery) and persuasion (the influence to adopt). Rogers writes 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.”

4. Mark Granovetter’s 1973 article, The Strength of Weak Ties offered an explanation of how information traverses social networks by positing that new information often comes to an individual from those in one’s network who are socially distant. Granovetter noted that while close friends tend to be homophilic, that is, alike in things like education, taste, opinions and views, close friends aren’t a good source for new information because there is a good chance the information is known already. And while homophily and strong tie relationships are important factors in persuasion, it is the weak tie who hold information not known to the strong tie individuals. The weak tie effect on sharing behavior on Facebook and other networks has been experimentally demonstrated. Another factor is that an individual tends to have more weak ties than strong ties.

5. Finally, from library science, we have theories and explanations for information seeking behaviors, and two of them often compared are search and browse. Searching – looking to find something specific – is more often than not focused, convergent, goal oriented and

---

systematic. Browse on the other hand, is less focused and used when specific information needs are not yet fully defined or understood. Browse then, compared to search, is divergent, dynamic and often undirected.

Successful browse depends on the seeker to recognize and make associations that serve individual needs for information. In the library literature, a term often applied to browse is serendipity, an unsought, unintended and/or unexpected discovery and/or learning experience that happens by accident and sagacity.

6. If a person knows the author and title of the book or has a very focused idea of what the book needs to be about, then conventional search strategies are often successful. But other strategies may be employed. A person may ask a friend or family member for a recommendation, or may visit a social website where readers hang out to browse favorite lists or author pages or may turn to blogs or other venues;

These are all cases where an individual may find and connect with people not in their social networks even weakly. Computer mediated communication channels make it relatively easy to find previously unknown individuals who may, through their posts and messages, trigger an information transfer to the browser, who in turn may create, forward or otherwise move the information across social networks where no prior measureable tie exists.
The 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.

I’ve called this phenomenon the serendipitous tie and defined it as:

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 slide shows a Venn diagram view of the classical view of information diffusion between networks via a shared weak tie relationship. Here, weak ties serve to bridge information flow between social networks.
This slide shows a view of the way information penetrates a social network either through serendipitous ties, in which information is exchanged between people who do not necessarily have weak tie affiliations, and directly from mass media influence.

Network researchers are aware that information comes to a network from external sources, but this is often something they often control for in research. The classical view of weak tie influence was confirmed in a study by Bakshy et al\textsuperscript{7} that looked at over 200 million wall posts across Facebook’s entire platform in order to study sharing patterns of strong tie and weak tie relationships, which, by the way, convincingly demonstrated Granovetter’s weak tie hypothesis.

In this study, as a method of control, they actually blocked some shared posts in order to determine a range of limits to the amount of information shared by individuals who did not get the information from strong or weak tie Facebook connections.

But I think that’s an important construct, and the serendipitous tie, as I’ve called it, is actually a pretty interesting phenomenon, in part because it appears that the serendipitous exchange of information emulates mass media effects as described by the two-step and diffusion of innovations theories I’ve mentioned.
Social Gatekeeping

• Finding, selecting, filtering and shaping information about a product, service or idea
• Making it available (or not) as a message accessible in a social communication channel.
• Web Presence
  • Computer-mediated Social Gatekeeping / Recommendation engines

So to summarize, 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.

The message is the unit of analysis, identified by a URL or some other kind of identifier.

Further, the more messages there are that are shared the greater the web presence of an author or book.

Once these messages are out there, not only people can find them, but applications and processes can find them, and these often form the basis of recommendation engines, which analyze, pool and extract social data as a marketing technology. So social gatekeeping can also be mediated by machine processes, and machine processes can connect people and their data serendipitously even without their explicit knowledge.
Slide 10 – The Traditional Publishing Chain Re-envisioned

Traditional Publishing Chain Re-envisioned

(An enlarged version of this graphic may be found in the Appendix). So now, let’s take a new look at the Traditional Publishing Chain. 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 a revised view of the traditional publishing model of the 20th century, shown on slide 10, 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 physical book is working its way through editing, design, printing and distribution channels, shown in green, information – or metadata – about the book is made available through mass media advertising and marketing and through a mass media network of reviewers and critics, shown in red.

The reading public might find it directly – for example, 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 the reader
might find it indirectly from a person of influence such as 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 20th 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.
Today’s Publishing Chain

(An enlarged version of this graphic may be found in the Appendix). However in the 21st century, as shown on slide 11, 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 channels such as Web sites, blogs, social network fan pages and other online social information constructs.

The expanded view of the publishing chain as it exists at the beginning of the 21st century shows that authors and publishers are 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.
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 which triggers the diffusion of information about the book through social networks and generally through the Web. Authors now have many potential pathways to enhance discoverability for potential readers.

The principle organizing research questions posed for this research 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? This is an important and fundamental question that needs to be addressed as the first step in testing the social gatekeeping framework. While this alone is insufficient to establish social gatekeeping as a robust gatekeeping theory extension, a negative result would serve to cast doubt on the framework’s viability.
Nature of the Study

1. Why eBooks?
2. Why Amazon
3. Random Sample of eBooks
   1. 484 ebooks from a pool of ~8000 released on Amazon in April of 2012
4. Popular Sample of eBooks
   1. 190 top selling / top downloaded ebooks on Amazon in April of 2012
5. Tracked and measured over 15 weeks in the summer of 2012

1. The research focuses on eBooks as an emergent form of literary production, now 20% or more of sales even among the major conglomerate publishers, and which by their very nature can only be acquired and read through some form of computer-mediated communication medium. This makes them ideal candidates for a study on the use of computer mediated communication by authors.

   Further, while many eBooks have print versions, many are digital only. Selecting eBooks as the object of research provides an opportunity to compare digital only versions with digital plus print versions to see if there are differences in the impact of author web presence on discoverability and sales.

2. A limitation of the study is that the results and interpretation are only strictly generalizable to books released and sold by Amazon. Amazon accounts for the majority of sales in the eBook market at an estimated 60% or even more according to some analysts.

   It’s true that 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 several potential sources reviewed, only Amazon provided both a search-browse function that could return a complete population of books in a non-biased return order and robust computer-based access to the eBook’s internal metadata. At this point in time, Amazon was 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.

3. So, the research is based on a random sample of eBooks drawn from the total population of 8,000 or so eBooks released on Amazon between March 31 and April 5, 2012 inclusive. The use of a true random sample drawn from a total population is a standard assumption of many statistical tests including Regression, which was the primary statistical tool used for the quantitative analysis that I’ll discuss shortly.

4. I also generated a list of the most popular eBooks from lists of the bestselling paid and most downloaded free eBooks on April 6, 2012. The second sample was used to provide a look at successful eBooks and compare these exemplars with books from the random sample.

There are some interesting differences between the two samples that you will see when I present some of the results. Note, though, 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 when interpreting the results I’ll present shortly.

5. Data about the eBooks were tracked and collected weekly for fifteen weeks throughout the summer of 2012 using custom software and largely automated methods. While that was
going on, information about publishers and author’s use of social media, along with other descriptive data, was collected manually through search techniques.

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, that is, the search engine hit count, 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 Amazon sales rank. Presence may also be reflected in counts of reader 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.

---

8Precision 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.
Slides 14 and 15 show the main data elements collected for analysis. The dependent variables designed to measure book web presence include search engine queries on

- ASIN which is the Amazon stock number assigned to each ebook, and searches on
- a specific quoted author title phrase,
- Amazon sales rank,
- Amazon review ratings and counts,
- and Offer price along with a few other miscellaneous data points.
Independent variables are examined to determine whether they predict differences in the measured dependent variables. Six methods of social media outreach were collected as the independent variables for author web presence:

- An Amazon author page, which an author may claim and post biographical information and links to other social sites,
- A Goodreads author page, which similarly provides a mechanism for authors to connect to readers on that social network,
- a Facebook page,
- a Twitter account,
- a Website and
- a Blog.

These were selected for review based on an early review of selected eBooks as being representative of the social media use by authors.
Methodology

- Phase I – Counts, Totals, Measures of Central Tendency
- Phase II – Multiple Regression to test main research questions and hypotheses
- Phase III – Qualitative look at selected titles for insight and future research.

The research naturally fell into three phases, which are interrelated and that, in total, provide evidence consistent with social gatekeeping from multiple perspectives.

Phase I included the collection, disaggregation and classification of the data collected. Statistics generated from phase I include totals and subtotals, plus some direct counts of sales that can be grouped by social media use and other categories such as mainstream published or self-published.

Because not all books in the Random or Popular samples were suitable for analysis of author social media outreach (for example, out of copyright classics, magazines released as eBooks and some other categories), the classification effort was also used to identify the subset of approximately 325 books used for analysis in phase II.

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 statistical tool used to determine the degree to which the independent variables could predict variance in the dependent variables. The
Social Gatekeeping framework 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 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, how authors may be leveraging the serendipitous tie, 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.
Phase I Highlights

- Re-publication of public domain books continues to represent a significant portion of title output.
- Self-published book counts (73%) exceed even Bowker’s estimates, possibly due to the number of self-published titles on Amazon without ISBN.
- Most books, and especially self-published books don’t sell well or at all, at least initially.
- Large numbers of eBooks are published without print equivalents (80%).
- There was an unexpected absence of enhanced ebooks and ebook apps (possibly an Amazon issue).
- Large numbers of short works, 1 – 20 pages or so, most self-published.
- Clear differences in “no sales” between authors who used social media and those who didn’t.

The research questions for Phase I are focused on sample description. I’ve also presented some of the data summarized in table format on the next slide. This slide shows some of the results I found particularly interesting.

- Re-publication of public domain books [the so-called reprint jungle] continues to represent a significant portion of title output, which reprises a 2008 study, on which I was a team member, on self-published books, as well as industry figures.
- Self-published book counts (73% of the total number of current authored titles) exceed even Bowker’s estimates, possibly due to the number of self-published titles on Amazon without ISBN.
- Most books, and especially self-published books don’t sell well or at all, at least initially.
- Large numbers of eBooks are published without print equivalents (80%).
- There was an unexpected absence of enhanced eBooks and ebook apps. Because these are highly promoted on the Apple platform, this may be a platform issue, but we don’t have a good idea even on Apple of the numbers of enhanced eBooks compared to the total number of eBooks. So this bears further investigation.

• In the random sample, but not the popular sample, I found several short works, 1 – 20 pages or so, most self-published. Some industry analysts think this is an emerging form of literary output; this confirms that at least to a degree, and it bears further research.

• There are dramatic differences in “no sales” – books offered for sale that reported no Amazon sales during the data collection window -- between authors who used social media and those who didn’t. Authors who didn’t use social media had almost double the number of no sales, as a percentage of total, as authors who used at least one form of social media outreach.

• See the chart No-Sales as a Function of Social Media Use slide 19. This is strongly consistent with the predictions of social gatekeeping theory.
## Selected Descriptive Statistics

(An enlarged version of this graphic may be found in the Appendix). Here are some selected descriptive statistics. You might want to pause playback if you’d like to study this in more detail.

<table>
<thead>
<tr>
<th>Category</th>
<th>N</th>
<th>n</th>
<th>Random Sample</th>
<th>N</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign Language</td>
<td>483</td>
<td>58</td>
<td>12.0%</td>
<td>190</td>
<td>0</td>
</tr>
<tr>
<td>Reprint Jungle</td>
<td>425</td>
<td>62</td>
<td>14.6%</td>
<td>190</td>
<td>8</td>
</tr>
<tr>
<td>Self-Published</td>
<td>360</td>
<td>262</td>
<td>73.0%</td>
<td>182</td>
<td>71</td>
</tr>
<tr>
<td>Other Editions (B&amp;N)</td>
<td>360</td>
<td>143</td>
<td>40.0%</td>
<td>182</td>
<td>137</td>
</tr>
<tr>
<td>Ebook only</td>
<td>422</td>
<td>156</td>
<td>36.0%</td>
<td>190</td>
<td>49</td>
</tr>
<tr>
<td>Self-published</td>
<td>336</td>
<td>234</td>
<td>70.0%</td>
<td>189</td>
<td>50</td>
</tr>
<tr>
<td>Description Lacking</td>
<td>423</td>
<td>43</td>
<td>10.0%</td>
<td>189</td>
<td>3</td>
</tr>
<tr>
<td>Review Included</td>
<td>580</td>
<td>14</td>
<td>3.7%</td>
<td>166</td>
<td>85</td>
</tr>
<tr>
<td>Kindle Loan Program</td>
<td>423</td>
<td>110</td>
<td>26.0%</td>
<td>189</td>
<td>71</td>
</tr>
<tr>
<td>Enhanced</td>
<td>423</td>
<td>0</td>
<td>0.0%</td>
<td>190</td>
<td>0</td>
</tr>
<tr>
<td>Optimized</td>
<td>423</td>
<td>13</td>
<td>3.0%</td>
<td>190</td>
<td>1</td>
</tr>
<tr>
<td>Erotica</td>
<td>423</td>
<td>36</td>
<td>8.5%</td>
<td>190</td>
<td>0</td>
</tr>
<tr>
<td>No Sales for the Period</td>
<td>322</td>
<td>76</td>
<td>24.0%</td>
<td>182</td>
<td>0</td>
</tr>
<tr>
<td>Self-Published - No Sales</td>
<td>76</td>
<td>68</td>
<td>89.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Self-Published - No Sales</td>
<td>235</td>
<td>68</td>
<td>29.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Price Manipulation - Lower</td>
<td>322</td>
<td>116</td>
<td>36.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Self-Published</td>
<td>116</td>
<td>77</td>
<td>66.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Price Manipulation - zero</td>
<td>322</td>
<td>36</td>
<td>11.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Self Published</td>
<td>36</td>
<td>32</td>
<td>49.0%</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
(An enlarged version of this graphic may be found in the Appendix). This chart summarizes Sales as a function of social media use.

<table>
<thead>
<tr>
<th></th>
<th>Used Social Media*</th>
<th>Didn't Use Social Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Sample, N=328</td>
<td>229</td>
<td>99</td>
</tr>
<tr>
<td>% that recorded no sales</td>
<td>19.0%</td>
<td>35.0%</td>
</tr>
<tr>
<td>Median Sales Rank**</td>
<td>328,465</td>
<td>499,751</td>
</tr>
<tr>
<td>% increase median number of books sold</td>
<td>36.0%</td>
<td></td>
</tr>
</tbody>
</table>

* # of types of media used: avg 2.2, median 2.0
** Lower Sales Rank = More Sales
Phase II - Regression and its Limitations

- Regression as a Statistical and Inferential Tool
- Limitations of Regression
  - Does not establish causality
  - Is subject to error under certain conditions – omitted variable bias, multicollinearity, reciprocal cause and effect.
  - Is a non-experimental method

Now, we move on to Phase II. As a statistical tool, regression is closely related to correlation, and the background math is very similar. But where a correlation compares two variables to see the correspondence between them, regression generally posits one or more variables as independent and one variable as dependent in order to see whether and to what extent a change in the independent variable might predict change in the dependent variable. So, regression tells us about the relationship between several independent or predictor variables and a dependent or criterion variable.

For this research, the dependent variable is one of the measurements of book web presence or sales, such as the search engine hit count or sales as estimated from sales rank. The independent variables are the six categories of social media use by authors, and they are tested as a group, or model, against each of the dependent variables. The purpose was to see whether and to what extent author web presence and social media outreach might predict discoverability and sales.
Multiple regression is one of the most commonly used statistical tools, but that means that it is also probably one of the most commonly misused statistical tools. Common pitfalls include assuming that because it tests the relationship of variables set up as independent and dependent, that it establishes causality, and that is simply not the case.

Also, there are certain conditions, some listed on the slide, where the regression may give misleading results. For example, if the regression is tuned by testing various combinations of independent variables in order to achieve maximum effect size, one variable may be given undue weight by leaving out a correlated variable. That’s omitted variable bias. And there are some other subtle traps.

And finally, regression is not an experimental method, which is considered the gold standard for hypothesis testing.
This slide shows key statistics you need to understand in order to follow regression, and I’ll explain what they mean and I’ll put them in context on the next slide when you see the results.

- \( R^2 \) is also called the coefficient of determination. It represents effect size, or the percent of variance observed in the dependent variable predicted by the independent variables.
- Model significance is measured by the F statistic and p value which together indicate the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming the null hypothesis is true. In this case, results were considered significant if these odds were no greater than 5%.
- Standardized Beta, or standardized coefficient, is the influence calculated for each of the individual predictor variables on the dependent variable, assuming the remainder of the independent variables are held constant, expressed in terms of standard deviations. Each Beta is also individually tested for significance. Only significant betas are usually reported.
- Regression indicators apply to the whole model including all the predictor variables tested; Betas for individual predictors may change as the model changes (predictor variables added to or deleted from the model).
(An enlarged version of this graphic may be found in the Appendix). The research questions all involved determining the various relationships between social media outreach and aspects of discoverability and sales. Social gatekeeping predicts that there will be a positive association between social media outreach on the one hand, and discoverability and sales on the other, and these are formally stated as hypotheses. Simply stated, the research tests whether authors who use social media outreach enjoy greater discovery and sales, and further, whether some kinds of social outreach might be more effective than others.

There is also a research question and hypotheses related to reviews.

In all, 14 regressions related to the main hypotheses were computed, and all were significant. The only model not reported for the Random and Popular Samples used Google search engine hits of Amazon Stock Numbers found on blog sites, and the hit counts were just too low to provide reliable statistics. The 12 remaining shown here were all significant at p < .05.

On slide 22 is the table for the Random sample, and as you can see, the $R^2$ values are in the low to moderate range, accounting from 18% to nearly 40% of the variance observed in each
of the dependent variables. Simply stated, that means that the social outreach accounted for between 18% and 40% of the difference observed in discoverability, sales and reviews. We might expect this since the model doesn’t include other things authors and publishers can do to increase discovery and sales such as mass media advertising, so I consider this a good result.

As you can see, a Goodreads author page had the best predictability of all the predictors. The way you might read these, if you look at the Amazon Reviews column, is that for each 1 standard deviation in the increase in author participation on Goodreads, you might expect a .38 (38%) standard deviation increase in the Amazon review count. Amazon recently bought Goodreads, and this number might explain why. I think it’s interesting that none of the remaining predictors were significant, including twitter, a blog or a web page.
(An enlarged version of this graphic may be found in the Appendix). In contrast, now looking at the results for the Popular sample, we see a shift. Goodreads is no longer much of a prediction factor, but Facebook ticks every dependent variable, and Web page ticks four of them. Overall, the effect size, that is, the $R^2$ values, is lower as are the Betas overall, compared to the Random sample.
Miscellaneous Regressions, DV = Sales

<table>
<thead>
<tr>
<th>Sales Predictor</th>
<th>R²</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available Print Version</td>
<td>0.252</td>
<td>0.141</td>
</tr>
<tr>
<td>Twitter # of tweets</td>
<td>0.078</td>
<td>0.243</td>
</tr>
<tr>
<td>Twitter # following</td>
<td>Not Sig.</td>
<td></td>
</tr>
<tr>
<td>Twitter # followed</td>
<td>Not Sig.</td>
<td></td>
</tr>
<tr>
<td>Facebook Friends/Likes</td>
<td>Not Sig.</td>
<td></td>
</tr>
<tr>
<td>Other books by author</td>
<td>Not Sig.</td>
<td></td>
</tr>
</tbody>
</table>

- Significant Predictors, Random Sample by Fiction: Amazon, Goodreads, Facebook
- Significant Predictors, Random Sample by Non-Fiction: Amazon

(An enlarged version of this graphic may be found in the Appendix). One thing you don’t want to do with a regression analysis is to throw a lot of independent variables at it. The more independent variables you include, the more cases you need to be comfortable with the numbers. When I designed the study, however, there were some other things I was curious about, so I ran selected independent variables against sales to see what the numbers look like.

I was not surprised to see that having a print version available might be associated with increased sales, but I was surprised to find out that when you look at social media use not as a dichotomized yes/no variable but actually look at numbers of interactions, only number of tweets was significant, and the number of twitter followers or the number of facebook friends was not. You might think that more followers would predict greater hits and sales, but that wasn’t the case here. Also not significant was whether an author had more than one available book. These are areas for further study.
Phase II Highlights

- Hypotheses supported by statistical results
- Effect size is low to moderate, which is to be expected.
- Emergence of Goodreads as most impactful predictor supports the social gatekeeping framework
  - 17 million avid readers who connect and post about books (23 million reader reviews).
- Facebook also about friends and family sharing, more so than blogs or websites.
- Different predictors for Popular sample
- Facebook and Web may play a more important role in driving sales once a title has come to the attention of a reader. Consistent with social gatekeeping and also multi-stage processes of adoption
- Caveats – causality not established; regression values indicate relative importance of but not absolute values of predictability; tests social media outreach but not social media generally

So what do these results mean?

- First, all statistical tests confirmed the basic hypotheses predicting a positive association between author social media outreach and book discoverability and sales. There was also a positive association between reader social reviews and eBook sales.
- The effect size is low to moderate, which I believe is to be expected
- The emergence of Goodreads as most impactful predictor in the Random sample supports the social gatekeeping framework:
  - Goodreads consists 17 million avid readers who connect and post about books with 23 million reader reviews. It has the highest concentration of individuals posting and sharing information about books of all the social media tested so it’s not surprising that of all the independent variables associated with social gatekeeping, Goodreads is the highest.
- Facebook is also about friend and family sharing, more so than blogs or websites.
- Different predictors emerged for the Popular sample, namely Facebook and Web pages, along with a lower effect size.
- So Facebook and the Web may play a more important role in driving sales once a title has come to the attention of a reader. This would be consistent with social gatekeeping and the theories on multi-stage processes of diffusion and adoption
- And finally, the caveats are:
  - causality is not established
• The regression values probably indicate relative importance of but not absolute values of predictability
• The regressions test social media outreach by authors but not social media generally, that is, although twitter (for example) wasn’t a significant predictor here, it doesn’t mean twitter may not be an important discovery tool – just not when used by authors.
Phase III Highlights

- Considerable variation in how social media was used by authors
- Raises questions of how to maximize social media effect
- Confirms the selection of dependent variables; only a few authors experimenting with alternatives, such as Pinterest, Stumbleupon, YouTube and LinkedIn
- Links to LibraryThing relatively infrequent
- Increase over time in social site sharing widgets, indicative of serendipitous tie
- Authors from the Popular sample are active participants in social media, perhaps effect as well as cause
- Some mainstream publishers expect authors to come to them with well-developed social strategies as a condition of publication

The research questions for Phase III were concerned with taking a qualitative look at some selected authors and titles for confirmation of the selection of variables, for additional insight into the results, and as inspiration for future research. In addition, I wanted to look for evidence of or ways to get more explicitly at the serendipitous tie since it is not really tested by either the phase I or phase II analyses.

The titles were selected on the basis of notes I made during phase I classification of anything I saw that I judged was either confirmatory or new and interesting. In all, about 35 titles from both samples were scrutinized and sites revisited nearly a year after the initial data collection.

I would summarize the most important impressions as follows.

- There is considerable variation in how social media was used by authors with some using it in rich and complex ways and others using it sparsely. This raises questions of how to maximize social media effect and how to evaluate both qualitatively and quantitatively the impact of social media use on discovery and sales.
• The review generally confirms the selection of dependent variables as appropriate and reasonably complete; only a few authors are experimenting with alternatives, such as Pinterest, Stumbleupon, YouTube and Linked-In, although those numbers appear to be increasing, especially among popular authors.

• I was surprised to see that links to LibraryThing were relatively infrequent, even though LibraryThing is a popular book-centric social network

• There was a rather dramatic apparent increase over time in social site sharing widgets, which can be used by authors to track sharing patterns by fans. This is the feature most indicative of the serendipitous tie and shows that authors understand the importance of encouraging sharing behavior by visitors to their sites. This suggests that future research on the serendipitous tie might be most fruitfully examined using link-tracking techniques since information shared without explicitly observable tie status leaves few traces otherwise.

• Authors from the Popular sample are active participants in social media, much more so than authors from the Random sample, which perhaps may be effect as well as cause.

• Some mainstream publishers expect authors to come to them with well-developed social strategies as a condition of publication. This suggests that publishing experts recognize the importance of social media as a catalyst for the diffusion of information about books through the reading public via social networks.
Slide 27 - Conclusions

Conclusions

• Findings are consistent with the social gatekeeping framework
• Sheds considerable light on the changing nature of the author-reader relationship
• This research is exploratory and designed to provide initial support for the social gatekeeping framework
• Opens the door to a research agenda to further and more explicitly test assumptions of social gatekeeping and the serendipitous tie
• Implications extend beyond book publishing

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 will 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.

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 information that could be mined from the data collected for this research that is beyond the scope of this research. So, there is more work to be done even on these samples.

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 as well as their role as a dependent variable measuring 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. The current involvement of libraries with eBooks is very unsettled.

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 to other disciplines and fields of study.
Figures and Tables

The Traditional Publishing Chain Re-conceptualized

Diagram of the traditional publishing chain, showing the process from author to reader, with various decision points and stakeholders involved in the publishing process.
Today's Publishing Chain

Author -> Manuscript

Accept?

Agent -> Represent

Accept?

Publisher -> Value Add

Distributor / Wholesaler

Sell?

Retailer

Buy / Read?

Reader

Publish / Distribute?

Marketing & Advertising

Review?

Traditional Media (TV / Newspapers / Magazines)

Share?

Social Network

Alternative Channels to Market (POD / eBook)
Research Questions, Phase I

- 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?

Descriptive Statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>Random Sample</th>
<th>Popular Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign Language</td>
<td>483</td>
<td>190</td>
</tr>
<tr>
<td>Reprint Jungle</td>
<td>425</td>
<td>190</td>
</tr>
<tr>
<td>Self-Published</td>
<td>360</td>
<td>182</td>
</tr>
<tr>
<td>Other Editions (B&amp;N)</td>
<td>360</td>
<td>182</td>
</tr>
<tr>
<td>Ebook only</td>
<td>422</td>
<td>190</td>
</tr>
<tr>
<td>Self-published</td>
<td>336</td>
<td>49</td>
</tr>
<tr>
<td>Description Lacking</td>
<td>423</td>
<td>189</td>
</tr>
<tr>
<td>Review Included</td>
<td>380</td>
<td>186</td>
</tr>
<tr>
<td>Kindle Loan Program</td>
<td>423</td>
<td>189</td>
</tr>
<tr>
<td>Enhanced</td>
<td>423</td>
<td>190</td>
</tr>
<tr>
<td>Optimized</td>
<td>423</td>
<td>190</td>
</tr>
<tr>
<td>Erotica</td>
<td>423</td>
<td>190</td>
</tr>
<tr>
<td>No Sales for the Period</td>
<td>322</td>
<td>182</td>
</tr>
<tr>
<td>Self-Published - No Sales</td>
<td>76</td>
<td>-</td>
</tr>
<tr>
<td>Self-Published - No Sales</td>
<td>235</td>
<td>-</td>
</tr>
<tr>
<td>Price Manipulation - Lower</td>
<td>322</td>
<td>-</td>
</tr>
<tr>
<td>Self-Published</td>
<td>116</td>
<td>-</td>
</tr>
<tr>
<td>Price Manipulation - zero</td>
<td>322</td>
<td>-</td>
</tr>
<tr>
<td>Self-Published</td>
<td>36</td>
<td>-</td>
</tr>
</tbody>
</table>
Notes on the preceding Descriptive Statistics Table:

- Reprint Jungle – out of copyright public domain works reprinted as new works
- Description lacking – book description lacking in the Amazon record
- Review Included – a trade or other review included in the Amazon description
- Enhanced / Optimized – application and/or best viewed on a tablet reader due to media or graphics
- Erotica – cataloged as erotica or adult as an Amazon subject classification (N.B. at the time of data collection, the eBook 50 Shades of Grey was not catalogued as erotica/adult by Amazon)
- Price Manipulation – authors experimenting with changing the price as a promotion, either lower or zero price (free download).
### No-Sales as a Function of Social Media Use

<table>
<thead>
<tr>
<th></th>
<th>Used Social Media*</th>
<th>Didn't Use Social Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Sample, N=328</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% that recorded no sales</td>
<td>19.0%</td>
<td>35.0%</td>
</tr>
<tr>
<td>Median Sales Rank**</td>
<td>328,465</td>
<td>499,751</td>
</tr>
<tr>
<td>% increase median number of books sold</td>
<td>36.0%</td>
<td></td>
</tr>
</tbody>
</table>

* # of types of media used: avg 2.2, median 2.0  
** Lower Sales Rank = More Sales
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?

- 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?

- 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

- 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.
Use of Social Media by Sample

% Author Web Presence

- Amazon Author
- Goodreads Page
- Blog
- Webpage
- Facebook
- Twitter

Random Sample vs. Popular Sample
Regression Results

Random Sample

<table>
<thead>
<tr>
<th>RANDOM SAMPLE</th>
<th>SIGNIFICANT PREDICTORS (p &lt; .05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Type</td>
</tr>
<tr>
<td>Google</td>
<td>ASIN</td>
</tr>
<tr>
<td>Google</td>
<td>Author/Title</td>
</tr>
<tr>
<td>Bing</td>
<td>ASIN</td>
</tr>
<tr>
<td>Bing</td>
<td>Author/Title</td>
</tr>
<tr>
<td>Amazon</td>
<td>Sales</td>
</tr>
<tr>
<td>Amazon</td>
<td>Reviews</td>
</tr>
</tbody>
</table>

Random Sample

<table>
<thead>
<tr>
<th>DV: Google ASIN</th>
<th>Google A/T</th>
<th>Bing ASIN</th>
<th>Bing AT</th>
<th>Amazon Sales</th>
<th>Amazon Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Squared</td>
<td>0.181</td>
<td>0.132</td>
<td>0.167</td>
<td>0.379</td>
<td>0.255</td>
</tr>
</tbody>
</table>

Significant Betas (p < .05):

| 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 |
### Popular Sample

<table>
<thead>
<tr>
<th>Popular Sample</th>
<th>Significant Predictors (p &lt; .05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amazon Author Page</td>
</tr>
<tr>
<td>Google</td>
<td>ASIN</td>
</tr>
<tr>
<td>Google</td>
<td>Author/Title</td>
</tr>
<tr>
<td>Bing</td>
<td>ASIN</td>
</tr>
<tr>
<td>Bing</td>
<td>Author/Title</td>
</tr>
<tr>
<td>Amazon</td>
<td>Sales</td>
</tr>
<tr>
<td>Amazon</td>
<td>Reviews</td>
</tr>
</tbody>
</table>

### Popular Sample

#### DV:

<table>
<thead>
<tr>
<th></th>
<th>Google ASIN</th>
<th>Google A/T</th>
<th>Bing ASIN</th>
<th>Bing AT</th>
<th>Amazon Sales</th>
<th>Amazon Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-Square</td>
<td>0.170</td>
<td>0.119</td>
<td>0.112</td>
<td>0.191</td>
<td>0.156</td>
<td>0.197</td>
</tr>
</tbody>
</table>

#### Significant Betas (p < .05):

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Author</td>
<td>0.164</td>
</tr>
<tr>
<td>Goodreads</td>
<td>0.100</td>
</tr>
<tr>
<td>Web Page</td>
<td>0.162 0.167</td>
</tr>
<tr>
<td>Facebook</td>
<td>0.243 0.185</td>
</tr>
</tbody>
</table>
Miscellaneous Regression Results

<table>
<thead>
<tr>
<th>Sales Predictor</th>
<th>$R^2$</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available Print Version</td>
<td>0.252</td>
<td>0.141</td>
</tr>
<tr>
<td>Twitter # of tweets</td>
<td>0.078</td>
<td>0.243</td>
</tr>
<tr>
<td>Twitter # following</td>
<td>Not Sig.</td>
<td></td>
</tr>
<tr>
<td>Twitter # followed</td>
<td>Not Sig.</td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td>Not Sig.</td>
<td></td>
</tr>
<tr>
<td>Friends/Likes</td>
<td>Not Sig.</td>
<td></td>
</tr>
<tr>
<td>Other books by author</td>
<td>Not Sig.</td>
<td></td>
</tr>
</tbody>
</table>

Phase III Research Questions

- 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?