Appendix A: Methodology

This appendix provides a more comprehensive overview of the data collection and analysis process. I used freely available packages including Gephi for network analysis, AntConc's corpus linguistics functions, and R for statistical analysis. Since Amazon's data are proprietary and several important sources are no longer publicly available, I can only share overviews of the data, but where possible I have indicated where the data can be accessed for replication purposes.

The research for *Four Shades* was conducted between 2013 and 2018. Over the five years, several of my e-readers and computers stopped functioning or were replaced, allowing for a variety of devices to be studied, including the following:

- Kindle Keyboard (3rd generation)
- MacBook Pro (2010)
- Kindle Touch
- Kindle 7
- iPad 3
- Sony Xperia E1
- MacBook Air (2015)
- Kindle Fire 5 (2016)
- HP 8300 Elite (Windows 2010; released in 2016)
- Kindle Cloud Reader (various devices)

Where possible, I noted the software versions used, but automated software upgrades on Apple and Android devices made it difficult to track exact versions, and change logs are not archived.

A Note on Archives and Sources

Working on a proprietary, commercially sensitive platform still in operation poses challenges for access. This is exacerbated in a case such as Amazon, which thrives on secrecy. In response to an interview question about not releasing sales figures for the Kindle, despite outselling predictions, Bezos stated, "It's tradition, mostly."[1] In place of institutional archives and interviews, one of the most reliable sources of information was the US Patent and Trademark Office's (USPTO) archives, which feature over seven thousand patents from Amazon employees. While the patents feature obsolete or theoretical technologies, they provide a conceptual framework for understanding the company's ideology when developing the Kindle. These are not a substitute for company archives and elite interviews, but patents can identify the company's strategic priorities more than documenting innovation. For example, Amazon's infamous patent filing for its "one-click" technology, discussed in further detail in chapter 2, reveals the aspects of Amazon's shopping technology that were considered important enough to protect.[2]

Capturing historical snapshots of Amazon's website, the Internet Archive was an invaluable source for *Four Shades*. Alexa, a company subsidiary, provided the Internet Archive with data about Amazon's web pages. The data set therefore represents the closest to an official open company archive, although the complexity of Amazon's website infrastructure ensures this is a limited snapshot, and coverage may be as patchy as less than 1 percent of all of Amazon's web pages. I supplement this evidence with data collated from MobileRead, a popular web-based forum for discussing the technical elements of ebooks, and a community that has tirelessly documented changes to software, hardware differences, and other otherwise ignored elements of the Kindle infrastructure. Finally, I used "View Source" on pertinent Amazon web pages to reveal the underlying data structures. I use this variety of sources to corroborate claims.

Chapter 2

Corpus Analysis (Table 2.1)

I used Brigham Young University's News on the Web (NOW) corpus—a cross section of online news articles published between 2010 and 2017—to assess

patterns in news organizations' portrayal of Amazon.[3] *Collocation* is the linguistic measure of how frequently two words appear in "proximity" to each other and how these words "subsist in the characteristic associations that the word participates in, alongside other words or structures with which [they] frequently co-[occur]."[4] For example, the meaning of "Apple" can be determined through its proximity to the word "iPad," "tree," or "Beatles." I conducted a collocate search of "Amazon" for any words appearing five words to either side. I removed words pertaining to the Amazon rain forest and clustered words into categories according to emergent themes.

Patent Analysis (Figures 2.1–2.2 and Table 2.2)

I discovered patents through the USPTO Patent Full-Text and Image Database (PatFT) search function. While a range of employees are named inventors for Amazon's patents, the parent company remains the "assignee." The exact name of the research and development company has shifted from "Amazon.com Inc" to "Amazon Technology," but a search for "(AN/"Amazon.com" OR AN/"Amazon Technologies")" catches all variations of the company's patent output. I classified the patents according to the filing date, which might be years before the patent was accepted. The analysis accounted for all granted patents as of February 5, 2018, the first batch of patents released for the month. The filing date is a more accurate marker of the company's position at the time of development than the eventual filing. For example, a subgroup at Amazon led by Janna Hamaker and Tom Killalea filed a series of patents related to book versioning and authority in 2010, but "Book version mapping" was not granted until December 2017, years after Amazon had moved on from developing new ebook technologies to focus on drones, other media, and cloud computing.[5]

Figure 2.1 was calculated directly from within the USPTO search by limiting the search to a year. Due to the lag in the USPTO granting applications and the ongoing review process, the data from 2013 onward indicate not a decline but rather that numerous patents filed since 2013 are still under review. Given the upward tick in filings, this is likely to exceed the output before 2013. Figure 2.2 was compiled by extracting the USPTO classifications for the patent applications and creating pairs where classifications collocated. I imported the data to Gephi to create a network. I manually identified the denser clusters and labeled them with the most common classification term and its relationship to the Amazon ecosystem.

I compiled table 2.2 through a keyword analysis of the abstract and claims of the textual corpus of Amazon's granted patents. The linguistic analysis of "keyness" focuses on words that appear more frequently in a corpus compared to a reference corpus. This technique is used to identify

the words that are more pertinent in one set of text compared to another. While "the" might appear more frequently in one body of texts than another, this does not translate to salience: the use of modal verbs is more common in instructional text than in poetry. Modal verbs should rank higher in terms of keyness, even if they appear less often than "the," "a," and other commonly used words. I downloaded the text directly from the USPTO PatFT full-text database and sorted patents into three-year intervals according to filing date. I calculated the keywords using three-year moving averages, so each set of keywords reveals the terms that appeared more frequently in a period compared to the preceding past years other than 1995–1997, which I compared to the 1998–2000 data sets. This highlights emergent trends in three-year periods, although it cannot be generalized to show that the terms in each list were a greater focus in that period than any other.

Chapter 3

Recommendations Network Analysis (Table 3.1)

The recommendation data were produced by a team led by Julian McAuley in 2014 in a project that required a large body of data from Amazon.[6] I used the subset of Kindle recommendations that contains a variety of metadata for 900,000 ebooks. Since Amazon limits recommendations to categories, it was possible to extract the "purchased with" and "also looked at" data for just the Kindle items without verifying that all books came from the Kindle Store. The size of the data set made a network visualization prohibitive and unlikely to yield interesting results. I instead gauged the scale and density of the network by calculating the volume of inbound and outbound links. After sorting them, I looked up the ASINs of the ten titles with the most inbound links.

Chapter 4

Since the Kindle creates different files according to the reading system configuration, the analysis of formats made full use of the diverse range of hardware I had at my disposal, including supplementary evidence from friends' and family's older devices. I collected a unique copy of my own ebooks through Amazon's "Download and Transfer via USB" option to access versions of files designed for my broken Kindles, including the first two Kindle Keyboards. Where possible, I loaded files onto the devices to capture any ancillary files included once the book is opened for the first

time. The iOS devices remained inaccessible owing to the obfuscation of folders without jailbreaking devices. Documentary evidence for the iPad file format therefore comes from secondary sources. I used a suite of tools including Hex Fiend (a so-called hex editor designed to read binary files) and Calibre, an ebook conversion tool, to analyze the files as both raw binary text and as they would appear if converted to EPUB. Due to the complexity of Amazon's format ecosystem, my analysis contains traces of Hex Fiend and Calibre. This comes with its own limitations, as Kirschenbaum warns that "both the emulator *and* the hex editor are programmatic computational environments applying some particular logic—a certain formal materiality—to the string of bits in question."[7] My approach in chapter 4 therefore mixes what Kirschenbaum has called "formal materiality," or "the relational attitudes by which [digital objects] are naturalized as a result of the procedural friction," and "forensic materiality," "the idea that no two things in the physical world are ever exactly alike."[8]

Chapter 5

Digitization of Titles from 1989 (Table 5.1)

The British National Bibliography (BNB), maintained by the British Library, offers records of every book published in the United Kingdom since 1950. I downloaded all records with a publication date of 1989. After cleaning the data, I had 30,940 titles published in 1989, including books reissued with a new ISBN. The resulting snapshot is arbitrary, but the sample size is sufficiently large to identify pertinent trends across genres and publishers. Since 1989 was a half decade before Amazon launched, all entries were intentionally uploaded as stock in its warehouses or via a third party, and therefore any evidence would demonstrate the company's commitment to maintaining a complete record of ISBNs rather than a direct import from Bowker's *Books in Print*. The main product page for each ISBN was manually examined and categorized according to each title's availability on the site during July 2017.

Amazon Charts

I collected data from the *New York Times* Best Sellers list and Amazon Charts between May and September 2017. The "Combined Print & E-Book" fiction and nonfiction lists were the closest equivalents to Amazon's "Most Sold." Since both lists are created using proprietary methodologies (including, for example, excluding certain genres), the data are not representative but instead reveal the ideological leanings of both data providers.

Chapter 6

Searching

All search results came from the Kindle for Mac 1.20.2 edition of John Milton's *Paradise Lost*. The latest version of Kindle for Mac (1.21) refactored the indexing system to discard all punctuation and spacing to consistently return 174 results regardless, removing some of idiosyncratic nuances visible in earlier versions of the indexing algorithm. This is not consistent with the indexing of Kindle for iPhone 6.3, however, which lists 420 results for "man." The inconsistency between reading systems is indicative of the discrepancies across the automated service paratext.

Chapter 7

"Social Reading" contains the largest data set, comprising over a million popular highlights and a selection of highlights from a further eight hundred titles. The volatility of the Kindle infrastructure during the five-year research period underpinning this book led to a degradation in the data's availability, including the removal of several titles in the eight hundred ebooks data set, and the slow erosion of the Kindle Popular Highlights website from the removal of the most popular list in 2016 to its closure in 2017.

The Kindle Popular Highlights Website

The primary website for viewing public highlights was started to allow users to access their highlights from an external location but grew into a valuable source of a slice of the data around the Kindle. I used two main sources to compile the data: (1) a list of the million most popular highlights based on static data from 2014, and (2) pages for each ebook title that contained a random selection of highlights pertaining to the book.

Eight Hundred Titles

The location data for the top ten most popular highlights are publicly available only through ebooks the user has either purchased or borrowed through Kindle Unlimited or Prime Reading. There is also no guarantee that a title, however popular, features a popular highlights list. I assembled the titles by sifting through the library of titles I had purchased for leisure or research in combination with free titles downloaded or borrowed through the Kindle Unlimited scheme during June 2017. The top ten popular highlights were extracted manually through a Kindle for Mac

application. No API exists for downloading popular highlights, as the data are available directly only through dedicated software.

I had initially hoped to equally represent the twenty-seven top-level categories Amazon offers for the Kindle Store, but categories such as "art and photography" are less likely to feature shared highlights than "fiction." Several issues emerged when attempting to find titles:

1. *Audience size:* Public domain books were the most likely to contain highlights, given their visibility as free books in comparison to the slurry of Kindle Unlimited. Smaller titles may occasionally contain single popular highlights or highlights that contradict information offered elsewhere. Since Kindle Unlimited offers limited access to best sellers, many of the books would be classified as midlist and not feature enough highlights.
2. *Genre expectations:* After sifting through more than one thousand titles to compile the corpus, I found clear differences between genres. For example, self-help guides marketed at women are more likely to be highlighted than books designed for pickup artists. Books that blur the lines between romance and erotica were more likely to contain highlights than those that had been classified as "adult." Since a substantial proportion of books did not feature highlights in an initial audit, I elected to focus on building a corpus from genres with a high probability of containing highlights.
3. *Comparability:* Amazon does not frequently update the publicly available public domain files, and as no data show how representative the highlights are, it is impossible to compare books based on their life span or publication. Every aspect had to be aggregated and indexed to allow for comparisons.

Due to these limitations, the collection overrepresents classic public domain titles and genres associated with the Kindle Direct Publishing boom, including self-help, romance and erotica, young adult, and fitness guides.

A collated list of the eight hundred titles and their ASINs (with a few exceptions) can be found online at sprowberry.com/800.xlsx.

Visible Highlights (Figure 7.3)

When Amazon introduced the "About the Book" sidebar, the feature introduced data about the total volume of highlights alongside the number of

unique selections. I compared these data with the sum of the top ten most popular highlights to calibrate how much material was unavailable through the ebook. The "About the Book" data do not feature dates when they were updated and, as with many aspects of Kindle Popular Highlights, can only offer an indication of the total volume. The analysis also reveals discrepancies between the data, as any result that moves closer to showing the complete set of highlights indicates an inconsistency in the numbers.

Highlighting Patterns

I created the heat map of a single quotation from Jane Austen's *Pride and Prejudice* (figure 7.10) by comparing highlights from the quotation from a range of editions that appeared in the 2015 list of most popular highlights. The data from Donna Tartt's *The Goldfinch* were extracted from the available data on the page dedicated to the title. I manually compiled a corpus of 2,239 publicly available highlights from *The Goldfinch* to assess variation between micro- and macro-level sampling. This represents only 1 percent of the total volume of highlights, but the sample is large enough to identify patterns that corroborate the data Amazon has aggregated about the most popular highlights. The sample was randomly generated by Amazon and therefore not representative but begins to identify patterns in the company's algorithmic curation of the popular highlights both within the reading system as indicated by patents and externally on the website.

Aggregate Reading Location (Figures 7.7–7.8)

I generated the boxplots in R from the internal highlight data from *Harry Potter* and *A Shade of Vampire* books published up until 2016. The boxplots show the distribution of the ten visible highlights in each book, showing the full distribution, median, and standard deviation.

Highlights as Text (Table 7.1)

I ran a word-frequency analysis of the 1.1 million highlights in the Kindle Popular Highlights database through AntConc, removing common articles and conjunctions to focus on nouns, adjectives, and verbs. The relative ranking was included to contextualize the position of the words within the wider corpus. I chose frequencies over a more sophisticated analysis such as keyness as an indication of words of interest, since a reference corpus would need to be representative of the books under discussion, many of which are still protected by copyright and remain unavailable for analysis.

Dictionary as Social Network

I scraped a set of comments on the Kindle Popular Highlights page for the *New Oxford Dictionary of American English*, second edition, in July 2016. The algorithmic mediation of the available comments only revealed data from the preceding week and year-old comments.

This is a section of doi:10.7551/mitpress/11985.001.0001

Four Shades of Gray

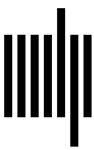
The Amazon Kindle Platform

By: Simon Peter Rowberry

Citation:

Four Shades of Gray: The Amazon Kindle Platform
By: Simon Peter Rowberry
DOI: 10.7551/mitpress/11985.001.0001
ISBN (electronic): 9780262369114
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Filosofia OT by Jen Jackowitz. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Rowberry, Simon Peter, author.
Title: Four shades of gray : the Amazon kindle platform / Simon Peter Rowberry.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Series: Platform studies | Includes bibliographical references and index.
Identifiers: LCCN 2021013279 | ISBN 9780262543507 (paperback)
Subjects: LCSH: Kindle (Electronic book reader) | Electronic book readers. | Electronic books.
Classification: LCC Z286.E43 R689 2022 | DDC 004.1675—dc23
LC record available at https://lccn.loc.gov/2021013279

10 9 8 7 6 5 4 3 2 1