DIGITAL PUBLISHING: THREE FUTURES (AND HOW TO GET THERE)♦

STEPHEN M. MAURER*

Abstract

The usual assumption that copyright rewards creativity is a fiction. In practice, most authors earn very little compared to their publishers. This article asks what services, if any, publishers supply to justify these payments. We argue that the only reasonable candidate is search, i.e. finding worthwhile titles among the million or so books written each year.

For most of the 20th Century, there was just one search technology: Human judgment. This led to a complex ecosystem of editors, bookstore owners, reviewers and other middlemen. The difference in the 21st Century is the emergence of a second technology—"Big Data"—that could make traditional methods obsolete. But in that case what new institutions will implement it? Depending on how Big Data evolves, we can anticipate three futures. In the first, the technology never advances much beyond its existing capabilities so that current institutions continue in something like their present form. We argue that this is already an improvement over mid-20th century publishing. At the same time, the advent of e-readers allows new forms of price discrimination that could significantly improve economic efficiency. Judges should reform the Second Circuit's Apple *decision to make this happen.*

More powerful "Big Data" technologies will force deeper

* Goldman School of Public Policy, University of California at Berkeley. The author thanks Diane Coyle, John Fingleton, Alex Karapetian, Sonja Garden, Will Grove, Fiona Scott Morton, Imke Reimers, Paul Seabright, Joel Waldvogel, and Ralph Winter. Special thanks to UC Berkeley's Center for Information Technology Research in the Interest of Society (CITRIS) for providing office space and other support for this project, The Toulouse School of Economics, The Institute for Advanced Studies at Toulouse and the J.-J. Laffonte Digital Chair for hosting me at their Workshop on the Future of the Book (Jan. 6, 2016) and the Economics of Intellectual Property Software and the Internet (Jan. 7–8, 2016). Above all, my thanks to the late Suzanne Scotchmer, who first encouraged me to study these problems.

changes. These will almost certainly start with massive vertical integration. Our second future analyzes the case where today's dominant on-line retailers continue expanding up and downstream. Despite obvious concerns, we argue that clearing away costly middlemen will almost certainly improve social welfare on net. We also consider an alternate future in which today's dominant publishers preempt retailers by creating an open search platform. Taking search outside traditional proprietary models can radically improve consumer welfare, but only if legislators are prepared to make correspondingly large adjustments to copyright law.

Finally, we ask which of our three futures is most likely. We argue that Big Data algorithms are inherently voracious, so that the future belongs to whichever institutions collect the biggest and most useful datasets. We identify the conditions under which proprietary solutions can outperform open source and vice versa. The article concludes by asking what judges and policymakers should do to create a level playing field so that the most efficient institutions really do emerge if and when technology makes them possible.

INTRODUCTION

Intellectual property (IP) rewards people who create new information. But what kind of information? And what people? Copyright theorists usually assume that IP provides an incentive for authors to create "content."[1] But this is untrue: In practically all times and places, most authors earn far less than their publishers.[2] This leads to a still deeper question: Given that copyright does so little to reward

[1] *Content*, WIKIPEDIA, https://en.wikipedia.org/wiki/Content_(media) (last visited Apr. 12, 2018).

[2] Complaints that author rewards are miniscule go back to Roman times. *See, e.g.*, MARTIAL, BOOK II 2.36 (complaining that author's wallet "doesn't notice" royalty payments). For more recent evidence, *see, e.g.*, John Eggen, *The Truth About Book Royalties*, EZINE ARTICLES (June 2, 2009), https://ezinearticles.com/?The-Truth-About-Book-Royalties&id=2424907 (only one in a thousand writers who contacts a literary agent gets published and makes any money). *See also*, Tom Shippey, *Book Review: 'A Bicycle Built for Brew' by Poul Anderson*, WALL STREET J. (Aug. 1, 2014), https://www.wsj.com/articles/book-review-a-bicycle-built-for-brew-by-poul-anderson-1406924607 ("In the 1950s there were only about five authors who made a living from sci-fi without needing a day job, and only one of them made a good living."). There is some evidence that authorship was briefly lucrative in the early 19th Century. *See*, Megan MacGarvie & Petra Moser, *Copyright and the Profitability of Authorship: Evidence from Payments to Writers in the Romantic Period* (Nat'l Bureau of Econ. Res., Working Paper No. 19521, 2013) (arguing that authors earned significantly more than working-class men); *see also*, George Orwell, *As I Please*, TRIBUNE (Mar. 3, 1944) (arguing that 19th Century literary markets were unusually lucrative), *found at* http://www.telelib.com/authors/O/OrwellGeorge/essay/tribune/index.html.

authorship, why have it at all?

One can imagine both small and large answers. The small response is to point out that publishers also contribute content, for example by telling authors how to revise their manuscripts, proofing texts, and creating cover art.[3] This line of argument is at least convenient, since it means that the familiar rationales for copyright can usually be redeployed with small adjustments. All the same, it is not very satisfying, since most of these tasks seem minor. Then too, they beg the question of why publishers need IP but authors do not.[4]

The larger answer is that content is just one type of information, and not necessarily the most valuable.[5] This article argues that the real purpose of copyright is to reward search, *i.e.* finding suitable titles among the one million or so titles published each year and matching them with readers.[6] This is a difficult and important problem. On the one hand, predicting what readers want is notoriously difficult so that the publishing industry frequently makes mistakes. On the other hand, the upside is enormous: Precisely because search is difficult, better predictions promise large benefits to readers.[7]

This article asks how today's search institutions evolved, how Big Data technologies can improve them, and what judges and policymakers should do to manage the transition. Section I describes the main technologies for predicting human taste. Section II describes the types and amounts of information that Big Data requires. We argue that, no matter how the technology evolves, its performance will normally be limited by supporting institutions' ability to assemble the data it feeds on. Section III recalls how today's institutions evolved to harvest human

[3] Common publishing activities include acquiring, developing, reworking, designing, producing, naming, manufacturing, packaging, pricing, introducing, marketing, warehousing, and selling books. Steven Piersanti, *The 10 Awful Truths about Book Publishing*, BK CONNECTION (Sept. 26, 2016), https://www.bkconnection.com/the-10-awful-truths-about-book-publishing.

[4] The most promising answer is that most authors write for reasons that have nothing to do with IP incentives, including vanity, reputation, or because employers like universities expect them to.

[5] Readers can, if they prefer, treat our search focus as an additional "lens" for analyzing copyright alongside more traditional theories. This viewpoint is likely to be most useful in fact patterns where search and content really are comparably important, so that policymakers must balance both goals.

[6] The US produces one million new book titles each year, of which 700,000 are self-published. An additional 13 million titles remain available from earlier years. This overhang is certain to grow given the explosion of new titles since 2010. Piersanti, *supra* note 3. I should add that the current article is deliberately centered on problems involving taste. This model is most relevant to general interest books in literature (both classic and potboiler) and popular non-fiction titles. The theory will be less useful for other titles including, for example, books for professionals (how to pass the general contractor's exam) and highly-specialized audiences (quantum field theory).

[7] The last big expansion in consumer choice increased reader benefits by nearly $1 billion in the year 2000. *See, e.g.*, Erik Brynjolfsson, Michael D. Smith & Yu (Jeffrey) Hu, *Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers*, 49 MANAGEMENT SCIENCE 1580, 1580 (2003) (estimating benefits attributable to rise of on-line booksellers in the early 21st Century).

judgment, stressing the many compromises and inefficiencies that have accrued in the process. This establishes our baseline for asking whether Big Data can do better. Succeeding sections argue that Big Data technologies will lead book markets into one of three futures. Section IV is the most conservative and argues that the future will look very much like the present. Here, the main feature is that digital book technologies have rewritten the copyright "bargain" by simultaneously eroding profits and inviting readers to do more search for themselves. We argue that this new tradeoff has made consumers better off, although legislative reforms might improve the balance still further. The revolution is nevertheless incomplete, mainly because the Second Circuit's *Apple* decision blocks publishers from implementing price discrimination that could make IP more efficient. Section V describes a second future in which further technology advances encourage today's on-line retailers to replace or marginalize other actors. We argue that this drastic streamlining is likely to be socially beneficial despite obvious monopoly concerns. Section VI presents our third and most ambitious future, describing how open institutions could supplant proprietary search altogether. This potentially offers large efficiency gains, but only if legislators are prepared to make correspondingly deep adjustments to copyright law. Section VII asks which of our three futures will prevail. We argue that markets will normally reward whichever set of institutions can best supply Big Data's ravenous appetite for information. Section VIII explores how real markets could nevertheless find themselves locked into other, inferior outcomes. Finally, Section IX discusses what judges and policymakers can do to create a level playing field that lets the most efficient institutions emerge if and when Big Data technologies are perfected.

I. THE TECHNOLOGIES

Innovation theorists usually focus on industries where the quality of new inventions can be captured in simple objective metrics like "horsepower" or "dollars saved." This simplifies the search problem to the point where institutional choices hardly matter. By comparison, the problem for cultural markets is far more difficult. Indeed, humans can and often do give wildly different quality estimates for the same goods. The result is that book publishers, in particular, devote much of their budgets to searching out titles that the public will want—and often fail. This section introduces the search problem and describes the two principal contender technologies for addressing it.

A. *Defining the Problem: Taste, Word-of-Mouth Markets, and Publishers*

The fact that people disagree about book quality is no more

surprising than that they have different personalities. What *is* strange is that they have so much difficulty anticipating how other humans will react. Indeed, even highly-paid experts routinely guess wrong about what the public wants. The point is commonly, if hyperbolically, made by quoting the familiar Hollywood aphorism that "nobody knows anything."[8] To some extent, the difficulty of predicting winners is inherent in human psychology.[9] But this is not the whole problem. Instead, much of the apparent unpredictability is injected afterwards through the peculiar dynamics of book markets.

The good news is that we understand these dynamics much better than we used to. Recent advances in both theory[10] and experiment[11] show that markets based on word-of-mouth recommendations are deeply capricious. The reason is luck-of-the-draw: If the first few generations of readers happen to be unreceptive, even good titles will be discarded. The cure, not surprisingly, is the Law of Large Numbers: Provided that the first few generations of readers are large enough, we can safely assume that they mirror the broader society. Beyond this, researchers report a sharp threshold where outcomes become dramatically more predictable.[12] This, however, implies the possibility of intervention, *i.e.* of organized efforts to identify and expose promising titles to the required critical mass of readers.[13] Historically, society has almost always entrusted this function to commercial publishers.[14]

Of course, knowing that intervention is possible is one thing. It is

---

[8] WILLIAM GOLDMAN, ADVENTURES IN THE SCREEN TRADE: A PERSONAL VIEW OF HOLLYWOOD AND SCREENWRITING 39 (1983).

[9] *See infra* Section I.B.

[10] Sociologists, computer scientists, and physicists have performed extensive simulations of how word-of-mouth recommendations propagate through networks. *See, e.g.*, F. Deschatres & D. Sornette, *Dynamics of Book Sales: Endogenous versus Exogenous Shocks in Complex Networks*, 72 PHYSICAL REV. 016112-1, 016112-5 (2005). For a technical description, *see* Duncan J. Watts & Peter Sherman Dodds, *Influentials, Networks, and Public Opinion Formation*, 34 J. CONSUMER RES. 441, 442–43.

[11] Salganik et al., exposed the same 48 songs to a series of artificial on-line markets, each containing thousands of listeners. They found that "hits" were highly random from one experiment to the next, although song quality made success somewhat more predictable. Matthew J. Salganik, Peter Sheridan Dodds & Duncan J. Watts, *Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market*, 311 SCI. 854, 854–856 (2006).

[12] Deschatres & Sornette, *supra* note 10, at 016112-6–016112-5.

[13] Assembling a critical mass of readers is expensive but straightforward. *See, e.g.*, ANITA ELBERSE, BLOCKBUSTERS: HIT-MAKING, RISK-TAKING, AND THE BIG BUSINESS OF ENTERTAINMENT 69 (2013) (quoting Disney executive Alan Horn: "You can buy an opening weekend . . . . It will be disappointing for you and for them, but you can get them in those seats.").

[14] Publishers' marketing efforts are supplemented by other commercial actors (bookstores, newspaper critics) and non-commercial recommenders like university faculty. The latter are particularly important for building titles into "classics" over the long term. Stephen M. Maurer, *The Economics of Memory: How Copyright Decides Which Books Do (And Don't) Become Classics*, 14 J. MARSHALL REV. INTELL. PROP. L. 520, 528–30 (2015).

quite another to say that publishers can pick winners well enough to justify the investment. Before the Digital Age, there was just one way to do this: Collect opinions from a large population of humans and then reconcile the differences to predict overall popularity. Looking back, commercial markets were a natural way to fund and organize this work. Even so, mixing strong IP rights with multiple middlemen created new problems that limited the practical benefits that better predictions should have delivered to readers and authors.

The problem, until recently, was that human judgment was the only game in town. The difference today is competition: Big Data has already demonstrated that it can make some predictions more cheaply, and could one day be more accurate as well.[15] But, in that case, today's institutions will soon be outdated and it is time to think about what should replace them.

B. *The Traditional Technology: Human Judgment*

Human judgment is notoriously unconscious and mysterious even to those who practice it. Still, it is important to say what we can. This section sets a baseline for asking whether Big Data can do better.

Sampling Strategies. The simplest theory denies that editorial choices are ever more than personal taste, or more precisely an opinion poll based on a single respondent. Beyond this, the idea that human judgment can predict what the broader public wants is mostly fictitious. Surprisingly, there is good anecdotal evidence for this view: Many authors and editors have said that that they do not even try to guess what audiences will like, but only publish what appeals to them personally.[16]

The fact that this improbable strategy works provides important information. In particular, it says that knowing one person's opinion (a) provides nearly complete information about significant numbers readers, but (b) provides almost no information about others. In what follows, we stylize these observations by imagining that American readers can be subdivided into multiple subpopulations or "reader groups."[17]

---

[15] *See* discussion *infra* Section I.C.

[16] *See, e.g.*, J.R.R. TOLKEIN, THE LORD OF THE RINGS 5 (2d ed. 1965) ("As a guide I had only my own feelings for what is appealing or moving, and for many the guide was inevitably often at fault."); Paul Elie, *Bound to Please*, WALL STREET J. (Aug. 3, 2013), https://www.wsj.com/articles/SB10001424127887324110404578625801325451188 ("This is literary publishing, and it's a business where your best guess about what might attract readers to give of their time is simply what you yourself are willing to spend time with."); Letter from Raymond Chandler to George Harmon Coxe, *in* SELECTED LETTERS OF RAYMOND CHANDLER 15–17 (Frank MacShane ed., 1981) ("I have assumed that there exists in the country a fairly large group of intelligent people . . . who like what I like.").

[17] We assume that each reader group is internally homogenous. Whether this is literally true is unimportant: Provided that we define enough groups, we can make any remaining intragroup

Theories of the Mind. The hypothesis that editors make decisions based on their own personal taste already explains how publishers can turn a profit.[18] Despite this, there are at least three reasons to think that editors can sometimes predict the reaction of reader groups they do not belong to:

> *Brain Science.* Biologists tell us that human brains have developed centers to simulate what others think. This "theory of the mind"[19] makes it plausible to think that editors can sometimes estimate tastes outside their own reader group.[20]
>
> *Best-Sellers.* Different reader groups often enjoy the same titles. This leads to "bestseller" strategies in which publishers seek out titles that attract more than one group at a time. At least in principle, predictions that a given title will appeal to multiple groups could be simpler and more reliable than deciding whether it would appeal to at least one group in the first place.[21]
>
> *Novelty.* Publishing thrives on novelty. But the most successful titles (*e.g. Harry Potter, Fifty Shades*) often attract readers who never bought books before and therefore cannot be identified from any existing sales data.[22] The fact that editors are aware of the wider society gives them a chance to do better.

If human judgment was limited to one's own reader group, we

differences as small as we like. It is worth noting that our "reader group" concept is closely related to "customer cluster truncation" technologies that try to save computation by assigning customers to pre-defined groups and then assuming that their tastes will mirror the group on average. In practice, this has not worked very well, presumably because they do not use enough categories. Clive Thompson, *If You Liked This, You're Sure to Love That*, N.Y. TIMES MAG. (Nov. 21, 2008), https://www.nytimes.com/2008/12/07/magazine/07letters-t-IFYOULIKEDTH_LETTERS.html.

[18] The economics of traditional print methods is surprisingly forgiving. In the mid-20th Century, books seldom needed to sell more than a thousand or so copies – one hundred-thousandth of the U.S. population – to break even. This meant that even editors with relatively uncommon tastes could usually find audiences. The math is even more favorable where some reader groups are bigger than others, so that editors from large groups have a built-in advantage.

[19] *Theory of the Mind*, WIKIPEDIA, https://en.wikipedia.org/wiki/Theory_of_mind (last visited Apr. 12, 2018).

[20] The inference is subject to the qualification that Nature sometimes hardwires behaviors that force members into individually irrational choices that benefit the species. *See, e.g.*, Eddie Dekel & Suzanne Scotchmer, *On the Evolution of Attitudes Towards Risk in Winner-Take-All-Games*, 87 J. OF ECON. THEORY 125, 140–142 (1999) (arguing that biology compels extreme risk-taking behaviors in males.).

[21] If so, the process remains highly imperfect. Hannah Furness, *Fewer than Half of Readers Finished Bestselling Novels*, THE TELEGRAPH (Dec. 10, 2014), https://www.telegraph.co.uk/news/shopping-and-consumer-news/11284934/Fewer-than-half-of-readers-finished-bestselling-novels.html (noting that most readers never finish the bestsellers they buy).

[22] *See, e.g.*, Kirsten Acuna, *BY THE NUMBERS: The '50 Shades of Grey' Phenomenon*, BUS. INSIDER (Sep. 4, 2013), http://www.businessinsider.com/50-shades-of-grey-by-the-numbers-2013-9.

would expect amateurs to pick winning titles just as often as professionals. But, in fact, professionals often seem to do better, even despite their errors.[23] In what follows, we assume that human judgment can indeed provide meaningful information outside one's own reader group.

C. *The Challenger: Big Data*

Human judgment is expensive and fallible. This makes it natural to ask whether advanced statistical and artificial intelligence methods can do better. Researchers began building "recommendation engines" to predict human taste in the 1990s.[24]

Recommendation Engines. In practice, the technology follows basic two strategies. The first, called "collaborative search," is summarized by Amazon's familiar advice that "People who liked X, also like Y."[25] Strikingly, this logic is completely agnostic: Indeed, it offers no underlying theory at all of why consumers might prefer one book to another.

The second strategy is called "content-based" search. It tries to assemble a theory, however simple-minded, of what consumers actually think. In operational terms, this means assigning "attributes" to both customers and products and then trying to match them.[26] For example, workers might be asked to say whether a particular title is "funny" or "sad." Conceptually, these judgments occupy a space midway between objective data and traditional editorial judgments. Like the former, they are cheap and highly replicable. But like the latter, they depend – however modestly – on unconscious thought processes that no computer

---

[23] William Goetzmann et al., *The Pricing of Soft and Hard Information: Economic Lessons from Screenplay Sales*, 37 J. CULTURAL ECON. 271, 297 (2013) (prices that studios pay for scripts have a "positive and significant" correlation to eventual box office revenue); Joel Waldvogel, *The Random Long Tail and the Golden Age of Television*, *in* INNOVATION POLICY AND THE ECONOMY 2016 (Shane Greenstein, Joshua Lerner & Scott Stern eds., 2017) ("While investors have some idea of which projects will do better than others, there is substantial uncertainty"); Clifton Fadiman, *The Reviewing Business*, *in* 260 A TREASURY OF AMERICAN WRITERS FROM HARPER'S MAG. 267 (1941) (Horace Knowles ed., 1985) (noting that the average critic "wouldn't hold his job long" unless his estimates were "appreciably more reliable than your dinner-table companion"); *see* Letter from Raymond Chandler, *supra* note 16, at 17 (" . . . I have never had any great respect for the ability of editors, publishers, play and picture producers to guess what the public will like. The record is all against them.").

[24] Thompson, *supra* note 17.

[25] *Id.*

[26] *See, e.g.*, Robyn M. Dawes, David Faust & Paul E. Meehl, *Clinical Versus Actuarial Judgment*, 243 SCIENCE 1668, 1671 (1989) ("Thus, for example, only the human observer may recognize a particular facial expression or mannerism (the float-like walk of certain schizophrenic patients) that has true predictive value."); Stuart J. Robbins et al., *The Variability of Crater Identification Among Expert and Community Crater Analysts*, 234 ICARUS 109 (2014) (crowd-sourced crater counts are comparable to those performed by experts); Peter G. Osorio & Michael J. Kurtz, *Automated Classification of Resolved Galaxies*, *in* DATA ANALYSIS IN ASTRONOMY III 122 (V. di Gesù et al. eds., 1989) ("Currently all galaxy types are obtained through visual inspection . . . ."); Miriam Cherry, *A Taxonomy of Virtual Work*, 45 GA. L. REV. 951 (2011).

can yet duplicate.[27] In practice, the most successful recommendation engines blend collaborative and content-based approaches together.

By definition, the future of these technologies is unknowable. We can, however, say something about where progress is likely to come from. Experience over the past quarter century suggests that most advances come from small, *ad hoc* experiments.[28] This suggests that researchers have no deep theory of search, so that further improvements will be incremental.

Still other advances in predictive power had less to do with algorithms than finding new types of information that probe individual taste more directly than traditional demographic data.[29] This notably included such information as the titles readers purchase or browse, the pages they read, and how quickly they read them.[30] The problem today is that most of the obvious strategies have now been implemented, so that it is hard to imagine companies inventing still better measures going forward.[31] Companies may, however, be able to expand their datasets over time. One of the biggest surprises in recent years is that e-books have plateaued at just under one-fourth of all titles sold.[32] It is reasonable to think that a truly convincing electronic substitute for paper[33] will eventually demolish this barrier, quadrupling the supply of data. After that, further expansions will depend on the extent to which publishers can develop incentives that convince readers to contribute

---

[27] Thompson, *supra* note 17.

[28] The fact that Netflix pays 1,000 employees to continuously tweak its algorithms typifies this strategy. Lara O'Reilly, *Netflix Lifted the Lid on How the Algorithm that Recommends You Titles to Watch Actually Works*, BUS. INSIDER (Feb. 26, 2016), http://www.businessinsider.com/how-the-netflix-recommendation-algorithm-works-2016-2.

[29] Netflix reports that traditional demographic data are far less predictive than knowing the specific movies that viewers have watched in the past. Thompson, *supra* note 17. Outside programmers who received access to selected Netflix data in the 2000s reported that the company's demographic information were too crude to be helpful. As one coder observed, "[t]here's little reason to think the other 40-year-old men on my block enjoy the same movies as I do." *Id.*

[30] Alexandra Alter, *Your E-Book is Reading You*, WALL STREET J. (June 29, 2012), https://www.wsj.com/articles/SB10001424052702304870304577490950051438304; *see* Joshua Tucker, *Making Sense of Kindle's Highlighting Feature*, SALON (Aug. 9, 2010), http://www.salon.com/2010/08/09/kindle_social_highlighting/; MICHAEL D. SMITH & RAHUL TELANG, STREAMING, SHARING, STEALING: BIG DATA AND THE FUTURE OF ENTERTAINMENT 329 (2016) (noting that Netflix tracks what consumers watch, when they watch it, what scenes they skip, and what scenes they watch repeatedly).

[31] *But see*, Robert Lee Hotz, *Songs Stick in Teens' Heads: Research Shows Hit Songs Activate Pleasure, Reward Centers in Adolescent Brains*, WALL STREET J. (June 13, 2011), https://www.wsj.com/articles/SB10001424052702303848104576381823644333598 (direct brain measurements predict "hits" better than respondents' own subjectively reported reactions).

[32] Piersanti, *supra* note 3 (reporting that e-book sales have declined slightly since 2012).

[33] A mature e-paper technology would ideally give readers physical books whose content could be repeatedly reprogrammed as necessary. Jason Heikenfeld et al., *A Critical Review of the Present and Future Prospects for Electronic Paper,* 19 J. OF THE SOC'Y OF INFO. DISPLAY 129 (2011); Iddo Genuth, *The Future of Electronic Paper*, THE FUTURE OF THINGS, http://thefutureofthings.com/3081-the-future-of-electronic-paper/ (last visited Apr. 12, 2018).

more and/or better information than they do today.

D. *Prospects*

For now, the case for Big Data remains, in the old Scottish phrase, "Not Proven." Logically, we can imagine the mature technology reaching three levels of capability. In the first, Big Data will never be much better than it is today. This will nevertheless be valuable to the extent that machines take over the simplest and most articulable rules of thumb. For example, readers who like one title in an established genre will enjoy other members as well. Beyond this, the technology could also search out statistical patterns that, once glimpsed, human experts can evaluate and judge to be true. In this modest future, Big Data will sometimes be cheaper but always less capable than traditional search.

The second possibility is that Big Data will find a way to replicate more complicated human judgments from simpler – but in some sense equally mysterious – "attributes" judgments. Here the early returns are promising. Indeed, standard personality tests already predict genre preferences (*e.g.* "poetry" vs. "crime" books) with high confidence[34] and can themselves be predicted from still simpler judgments including Facebook "Likes."[35]

The final possibility is that Big Data will eventually make better predictions than humans. This would be less surprising than it sounds. Psychologists have known since the 1980s that machines *always* out-predict humans for problems that can be reduced to actuarial tables.[36] More recent evidence has extended the argument by showing that Big Data algorithms sometimes access non-human "categorizations . . . so obscure that [the creators] cannot see the reasoning behind them."[37]

---

[34] Ivan Candator, Ignacio Fernandez-Tobias & Alejandro Bellogin, *Relating Personality Types with User Preferences in Multiple Entertainment Domains*, UMAP WORKSHOPS (2013); Nicola Schutte & John Malouff, *University Student Reading Preferences in Relation to the Big Five Personality Dimensions*, 25 READING PSYCHOL. 273, 279 (2004); William Tirre & Sharvari Dixit, *Reading Interests: Their Dimensionality and Correlation with Personality and Cognitive Factors*, 18 ELSEVIER 731 (1995); Peter J. Rentfrow, Lewis R. Goldberg & Ran Zilca, *Listening, Watching, and Reading: The Structure and Correlates of Entertainment Preferences*, 79 J. PERS. 223 (2011).

[35] Wu Youyou, Michal Kosinski & David Stillwell, *Computer-Based Personality Judgments Are More Accurate Than Those Made by Humans*, 112 PROC. OF THE NAT'L ACAD. OF SCI. 1036, 1036–1039 (2015). Facebook "Likes" can reportedly predict the results of self-administered personality tests better than the subject's own friends. *Id.*

[36] DANIEL KAHNEMAN, THINKING, FAST AND SLOW 234-44 (2011); Dawes et al., *supra* note 26 (finding that actuarial formulas invariably out-perform human judgment in predicting patient health and human behaviors); William M. Grove, *Clinical Versus Statistical Prediction: The Contribution of Paul E. Meehl*, 61 J. OF CLINICAL PSYCHOL. 1233, 1240 (2005) (confirming the superiority of actuarial methods).

[37] Thompson, *supra* note 17. Thompson adds that "[p]ossibly the algorithms are finding connections so deep and subconscious that customers themselves wouldn't even recognize them." For example, music buyers who enjoy classical composers disproportionately like the Beatles, and moviegoers who enjoy the family drama "Pay It Forward" also like sci-fi movie "I, Robot."

This last effect will become much more significant if, as many artificial intelligence experts claim, machine learning improves dramatically once computing resources cross a certain threshold. This may be happening already.[38]

E. *The Merger Problem*

Previous sections have discussed Big Data and human judgment as if they were binary alternatives. In practice, it seems more likely that each can improve the other.[39] But in that case, society must find some way to merge the two. There are three possibilities:

> *Machines-on-Top*. The simplest way to merge human judgment with machine predictions is to reduce the former to numerical scores that Big Data can ingest like any other variable.[40] The problem, as we have seen, is that humans could be shrewder than any algorithm.
>
> *Humans-on-Top*. The alternative is to let humans review Big Data estimates using the same unconscious processes that define "taste" more generally. In practice, even highly numerate firms like Netflix still make decisions this way.[41] At the same time, the movie industry offers a cautionary example of how creative disputes can waste resources, encourage stultifying compromises, and make aesthetic choices hostage to office politics.[42] These organizational impediments are still more complicated in the digital era, where

Id.; *see also*, Andy Kessler, *Bad Intelligence Behind the Wheel*, WALL STREET J. (Apr. 24, 2017), https://www.wsj.com/articles/bad-intelligence-behind-the-wheel-1492983234 (explaining how artificial intelligence discovers and exploits patterns that humans cannot sense or understand). *See also*, W. Kip Viscusi & Richard J. Zeckhauser, *The Perception and Valuation of the Risks of Climate Change: A Rational and Behavioral Blend* 12–15 (Nat'l Bureau of Econ. Research, Working Paper No. 11863, 2005) (people who worry about global warming tend to fear the risk of a heart attack); Michal Kosinski, David Stillwell & Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROC. OF THE NAT'L ACAD. OF SCI. 5802, 5804 (2013) (finding that statistical correlation between people who like "Curly Fries" and high intelligence).

[38] *See also*, Kessler, *supra* note 37 (arguing that increases in processing power have rapidly expanded artificial intelligence capacity since 2015).

[39] Grove, *supra* note 36 at 1237–38 some psychologists believe that humans can outperform actuarial tables in the special case where they receive access to the statistical prediction and are allowed to overrule it). Dawes et al., *supra* note 26, at 1671 ("If clinicians were more conservative in overriding actuarial conclusions they might gain an advantage, but this conjecture remains to be studied adequately."); Leslie Scism, *Insurance: Where Humans Still Rule Over Machines*, WALL STREET J. (May 24, 2017), https://www.wsj.com/articles/insurance-a-place-where-humans-not-machines-rule-1495549740 (quoting AIG Chief Underwriter Mahdu Tadikonda: "The models by themselves are not perfect." So that when an underwriter "turns off his or her brain, we're done.").

[40] Grove, *supra* note 36.

[41] O'Reilly, *supra* note 28.

[42] *See, e.g.*, GLENN FRANKEL, HIGH NOON: THE HOLLYWOOD BLACKLIST AND THE MAKING OF AN AMERICAN CLASSIC (2017); CHAOS ON THE BRIDGE (Vision Films 2014).

> many corporate cultures help software engineers win arguments with "book people" even when they shouldn't.[43]
>
> *Market Transactions.* Instead of trying to merge human judgment and Big Data within a single organization, publishers could instead take the Coasian path of committing the problem to arm's length market transactions. Since the 18th Century, publishers have let readers sort out quality following an initial ad campaign.[44] On the other hand, sellers may sometimes know more – or at least devote more effort to search – than readers themselves. In this situation, it might make sense to let retailers pay publishers to re-rank individual titles above recommendation engines' initial estimates.[45]

For now, we cannot be sure which solution will turn out to be efficient. However, it is worth noting that the problem is not very different from merging multiple human judgments together. Here, publishing has relied on arms-length Coasian transactions for more than 2,000 years.[46] We should not be surprised if this pattern persists into the Big Data era.

II. FEEDING BIG DATA

So far we have compared Big Data and human judgment on the assumption that information is abundant and free. This Section

---

[43] George Packer, *Cheap Words*, NEW YORKER (Feb. 17, 2014), https://www.newyorker.com/magazine/2014/02/17/cheap-words (describing Amazon's hacker culture); SMITH & TELANG, *supra* note 30, at 324, 327–28 (arguing that the publishing, music, and motion-picture industries have cultural barriers against data-driven analytics). One partial work-around is to create an independent division with a separate and distinct culture. Amazon has gone down this path by insulating its Hollywood-based Amazon Studios from the rest of its famously cost-cutting culture. Joe Flinty, Ben Fritz, and Laura Stevens, *Roy Price's Alleged Trail of Drinking and Sexual Harassment Challenges Amazon's Culture*, WALL STREET J. (Nov. 7, 2017), https://www.wsj.com/articles/roy-prices-alleged-trail-of-drinking-and-sexual-harassment-challenges-amazons-culture-1509986006. This is probably just a palliative, since it only replaces tyranny-by-programmers with what could be an equally dysfunctional tyranny-by-editors. The great advantage of a genuinely independent entity is that survival depends on the market, forcing culture to develop in whatever direction performs the most efficient mergers.

[44] *See* RICHARD B. SHER, THE ENLIGHTENMENT AND THE BOOK 361–69 (2006) (describing 18th Century marketing techniques); *see* Paul Elie, *supra* note 16 (modern publishers rely on "big talkers and social networks" far more than marketing).

[45] Auctioning search rank would raise obvious concerns given US and European regulators' recent claims that Google manipulated search results to make its own services more prominent. Significantly, US investigators ultimately found the practice acceptable, arguing that any negative impact on Google's competitors was "incidental" to improving the company's search results. *Google's Search Practices: In the Matter of Google Inc.*, FTC File Number 111-0163 (Jan. 3, 2013) (statement of the F.T.C.). A system that openly auctioned bid rank should be *a fortiori* acceptable under this standard. The European Commission has taken a harder line, although the matter is still under appeal. *See* Press Release, European Commission, Antitrust: Commission Fines Google € 2.42 billion for abusing dominance as search engine by giving illegal advantage to its own comparison shopping service (June 27, 2017). Probably the best argument for an auction system is that it is inherently honest. Publishers who pay to increase the visibility of titles that readers dislike will reliably lose their investments.

[46] *See* discussion *infra* Section III.A.

introduces economics and scarcity. We begin by describing the main types of data in economic terms. We then ask how Big Data's voracious appetite for information compares to what readers can plausibly supply.

A. *New Types of Data*

Modern recommendation engines extract predictions from many different types of data. These can usefully be grouped within three broad categories.

Objective Data. Some data are so straightforward that machines can collect them without human intervention. This notably includes reading or shopping habits recorded in the course of ordinary business operations. Other, older types of objective data include demographic information like customer zip codes and income. The former cost little or nothing to collect while the latter can usually be purchased from government or commercial vendors[47] or collected *gratis* from readers.

The prospect of free data is hugely attractive. That said, the supply is fixed by customers' existing shopping behaviors. We have argued that technology will provide a one-time jump with arrival of e-paper. Beyond that, further expansion depends on persuading readers to engage in more (or at least different) shopping behaviors than they do today. Here the obvious method is to offer subsidies and discounts. But these will only be sustainable if improved predictions generate enough additional sales to pay for themselves. This is unlikely in an era where piracy limits potential profits. Small subsidies could, however, shift readers' choices from popular titles to similar if currently more obscure alternatives. This would allow firms to trade mostly redundant data about how, say, readers respond to bestsellers for more interesting information about new titles or consumers themselves.

Simple Judgments. We have said that many recommendation engines depend on assigning simple attributes to books and readers.[48] These seldom cost more than a few seconds' effort by human coders and are highly replicable from one worker to the next. That said, no computer can yet duplicate them. This closely resembles the editor judgments that have supported search since Roman times.

The fact that simple judgments are replicable means that they can be purchased on the open market. Many firms do this by paying for piecework on-line,[49] while others hire full-time employees.[50] Still other

[47] Thompson, *supra* note 17 (noting that Netflix CEO Reed Hastings says that even though Netflix possesses a large stock of demographic data, it doesn't use them to generate movie recommendations).

[48] The vast majority of "Human Intelligence Tasks" or "HITs" on Amazon's Mechanical Turk site fit this description. *See Human intelligence through an API*, AMAZON MECHANICAL TURK, https://www.mturk.com/ (last visited Apr. 12, 2018).

[49] Piece-work is easily purchased from on-line sites. *See id.*

[50] *See* Thompson, *supra* note 17 (Pandora has fifty employees who listen to songs and then tag

data are donated. Theory teaches that donations are easiest when the minimum useful contribution is simple or granular.[51] This neatly explains why users have donated tens of millions of "Likes" on Facebook.[52] The deeper mystery of why some people donate hundreds of judgments at a time.[53] One possible explanation is that firms often ask consumers to describe themselves, which could be its own reward. However, this cannot be the whole story, since it fails to explain why volunteers also donate product descriptions[54] and other plainly industrial tasks.

Complex Judgments. This category includes the holistic and largely unconscious processes by which human readers assess books. The main innovation for Big Data is that it tends to harvest this information from lay people instead of expensive experts.

Complex judgments almost always cost more to acquire than data from "simple judgment" or "objective data" categories. The reason is that readers must usually invest hours of reading to arrive at a useful opinion. Costs fall dramatically, however, when respondents have read a particular title already, so that only the reporting costs remain. This latter burden typically varies from a few seconds (filling out a five-star rating) to perhaps an hour (drafting an essay).

The surprise, once again, is how many complex judgments are donated. Amazon's Goodreads site[55] generates twice as many book reviews as commercial sources.[56] The main drawback is that coverage is skewed to popular titles and a tiny (and probably atypical) minority of readers.[57] The fact that these judgments often disagree tells us that they say as much about readers as the underlying title.

---

them with descriptors like "upbeat," "minor key", or "prominent vocal harmonies."); Alex Iskold, *The Art, Science, and Business of Recommendation Engines*, READWRITE (Jan. 16, 2017), https://readwrite.com/2007/01/16/recommendation_engines/; Thompson, *supra* note 17 (Netflix has considered hiring cinephiles to write tags for all 100,000 movie in its library).

[51] Yochai Benkler, *Coase's Penguin or Linux and the Nature of the Firm*, 112 YALE LAW J. 369 (2002).

[52] Candator et al., *supra* note 34 (220,000 Facebook users had self-reported 46 million "Likes" as of 2013).

[53] For example, 3.1 million volunteers have completed standard personality tests that include hundreds of attributes judgments. *Id.* Typical examples of attribute questions include whether a person "makes friends easily" or has "a vivid imagination." *See The 300-Question Personality Test*, TRUITY, https://www.truity.com/test/300-question-personality-test (last visited Apr. 12, 2018).

[54] *See* Alex Iskold, *supra* note 50 (discussing how Del.icio.uslets users annotate products).

[55] GOODREADS, http://www.goodreads.com/ (last visited Apr. 12, 2018).

[56] *See* Joel Waldfogel, *Copyright and Technological Change, and the Quality of New Products: Evidence from Recorded Music since Napster*, 55 J. L. & ECON. 715, 735−39 (2012).

[57] Eric T. Anderson and Duncan I. Simester, *Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception*, 51 J. MARKETING RESEARCH 3 (2014), *found at* http://web.mit.edu/simester/Public/Papers/Deceptive_Reviews.pdf (estimating that 1.5% of the firm's customers wrote reviews); *What Percentage of Buyers Write Reviews on Amazon?*, QUORA, http://www.quora.com/What-percentage-of-buyers-write-reviews-on-Amazon (last visited Apr. 12, 2018) (estimating that 0.5-5% of Amazon buyers write reviews).

B. *Feeding the Beast: How Much Data Do We Need?*

Technology is not the whole story. Performance also depends on how much data readers and publishers can supply to feed Big Data's algorithms. The natural guess is that we expect opinion data to be cheap for heavy readers, burdensome for mid-rank consumers, and prohibitive for everyone else. The present section makes this intuition more precise by posing a thought experiment: What is the minimum amount of information that a *perfect* Big Data technology would need to make predictions?

Estimating The Number of Reader Groups. Assume that Big Data reaches the point where it can make perfect predictions with perfect efficiency. How much information will it need? We have already argued that "prediction" is identical to answering the question "which reader group does each customer belong to?" But this implies that the amount of information depends on the number of reader groups. Granted that this last figure is poorly known, we can at least narrow the uncertainty.

The lower bound is easiest. We have already said that real life publishers often make mistakes. In principle, they could avoid this by hiring enough editors to mirror the entire population. But, of course, they have not done so. This shows that the total number of reading groups must be at least as large as their editorial staffs.[58] We conclude that the total number reader groups is at least one hundred.[59]

Setting an upper bound is harder. Plainly, editors could not rely on personal taste as a guide to commercial viability if the number of reader groups was more than 100,000 or so.[60] A better, if somewhat anecdotal argument is that most of us know someone – either personally, on-line, or through the media – whose tastes reliably predict our own regardless of genre or title. For this to happen, the number of reader groups must be significantly less than one thousand.[61] Putting these observations

[58] For example, Simon & Schuster employs twenty-one full-time editors in its main office along with slightly smaller groups at each of its twelve adult reader imprints. *Our Team*, SIMON & SCHUSTER, http://simonandschusterpublishing.com/simonandschuster/our-team.html (last visited Apr. 12, 2018) (listing employees whose titles include the word "editor" or editorial"); for example, *Editorial Team*, TOUCHSTONE, http://www.simonandschusterpublishing.com/touchstone/about.html (last visited Apr. 12, 2018) (Simon & Schuster subsidiary imprint).

[59] The argument is probably conservative since editors have many tasks besides evaluating titles. This is partly offset by the fact that each publishing house serves slightly different reader groups compared to its rivals. For reasons that appear below, these refinements would not materially change our argument.

[60] We have said that mid-20th Century publishers needed to sell roughly 1,000 copies to break even. Dividing the American population into 100,000 groups would push the average reading group below this threshold. More refined estimates would take account of the fact that, for example, many Americans read no books at all, some reader groups are bigger than others, publishers are unlikely to sell to every group member, and/or some titles appeal across multiple groups.

[61] Sociologists estimate that the average American knows 600 people. *See* Andrew Gelman, *The*

together, we conservatively estimate that there are between 100 and 1,000 reader groups across the country.

Estimating Effort. This order-of-magnitude uncertainty might not seem very helpful. But what we really care about is economics, or, more specifically, effort. Here the arithmetic is more favorable. We therefore ask (a) how many questions a maximally efficient Big Data algorithm would need to ask, and (b) how much effort would be required to answer each question.

One preliminary difficulty is that some questions are more complicated than others. Following Shannon,[62] we start from the observation that all information is reducible to a common currency of "yes-no" questions or "bits." But in that case, a perfectly efficient Big Data technology would still need to ask at least seven binary opinions to assign readers to one of 100 reading groups[63] and ten opinions for 1,000 groups.[64] This range is so narrow that the precise number barely changes our estimate.

Finally, we need to know how much effort each answer requires. Here, the most natural guess is that respondents would have to read one book per query. While some abridgement is surely possible, this gambit seems limited on technical grounds.[65] More importantly, we have said that it is cheaper for readers to provide opinions about books they have already read. This implies that the publishing industry will normally collect more information by letting readers report whatever books they

Average American Knows How Many People?, N.Y. TIMES (Feb. 18, 2013), http://www.nytimes.com/2013/02/19/science/the-average-american-knows-how-many-people.html. The number of people who repeatedly give others advice about which books to read is almost certainly smaller, even accounting for virtual contacts with book reviewers in the traditional press and on-line media.

62 JIMMY SONI & ROB GOODMAN, A MIND AT PLAY: HOW CLAUDE SHANNON INVENTED THE INFORMATION AGE (2007); JOHN R. PIERCE, AN INTRODUCTION TO INFORMATION THEORY: SYMBOLS, SIGNALS AND NOISE (1980).

63 A survey that asks N binary questions can be answered in 2^N distinct ways. Thus, a survey that asks two questions can identify a maximum of $2^2 = 4$ distinct groups, while a survey that asks seven questions can diagnose $2^7 = 128$ groups. In principle we could reduce the required number of titles by replacing binary questions with more nuanced five-point "star" rankings. This, however, is an illusion unless readers actually use the additional freedom. Inpractice, most real star rankings clump near the top of the scale. *See e.g.*, Max Woolf, *A Statistical Analysis of 1.2 Million Amazon Reviews*, MINIMAXIR (June 17, 2014), http://minimaxir.com/2014/06/reviewing-reviews/ (more than half of all electronic product reviews award five stars). Systems which monitor how readers consume books at a page-by-page level might conceivably work better, but would run into the related objection that it is hard to imagine any one title diverse enough to fully diagnose taste.

64 $2^{10} = 1024$. Diagnosing the unfavorable and unlikely case of 100,000 groups would require 17 questions, since $2^{17} = 131,072$.

65 Some abridgment must be feasible: After all, most of us would reliably reach the same opinion even if a book's last page was missing. That said, the savings are limited. Anecdotally at least, readers often like novels less than they thought they would, and sometimes change their minds completely after a promising start. The deeper problem is that a novel is, or should be, a unified entity. If it was possible to replicate the experience in a shorter format, American authors would sell more short stories than they actually do.

happen to have read already.[66] Assuming optimistically that the resulting randomness degrades efficiency by a factor of a few, we conclude that *a perfect Big Data technology would need to collect reader opinions for about fifty titles to achieve its full predictive potential.* Real Big Data technologies will likely do worse, and cannot do better.[67]

C. *Implications*

Our analysis suggests that mature Big Data technologies must obtain consumers' opinions of roughly fifty titles. This immediately implies a three-class society. On the one hand, the bar is not particularly challenging for the top decile of American readers, who consume more than fifty titles per year or 500 per decade.[68] In this special case, a mature Big Data technology really could supplant human editors. Alternatively, it could be that human editors can make some predictions that Big Data cannot. In that case, elite readers might value better predictions enough so that publishers could pay human editors to improve Big Data's recommendations still further.

By comparison, Big Data would be a near thing for average readers who consume just five books each year.[69] Here the best predictions will come from whichever institutions collect the biggest and best blends of information. We argue in Section VII that open institutions might or might not outperform proprietary models depending on the specific types of information that a mature Big Data technology would need.

Finally, Big Data predictions will always be limited for the bottom

---

[66] Self-selected titles are bound to elicit more redundant information than an optimized questionnaire. This is particularly likely since readers often reduce risk by deliberately revisiting familiar authors and genres, or else consume bestsellers that have been engineered to please as many reader groups as possible. We should also expect market distortions. For example, the fact that readers get bored with genre fiction implies that quality scores depend on the order in which books are consumed. But in that case, popular books that readers find and read first will receive different (and probably better) scores than more obscure titles. Market effects are particularly insistent where people read books mainly as an excuse to socialize so that even bad books can become bestsellers. *See* ELBERSE, *supra* note 11.

[67] In principle, publishers could deploy a shorter questionnaire that deliberately left some reader groups undiagnosed. This could improve the cost-benefit ratio under special fact patterns, for example where some reader groups contain many more members than others.

[68] *Poll: 28 Percent of Americans Have Not Read a Book in the Past Year*, HUFFINGTON POST (Oct. 7 2013), https://www.huffingtonpost.com/2013/10/07/american-read-book-poll_n_4045937.html (out of 1,000 adults surveyed, 8 percent reported that they read more than 50 books in the past year). It is hard to see how readers could remember, much less give meaningful opinions for titles they read much more than a year ago. Indeed, many respondents will not have lived long enough to do this even in principle. The problem is still worse if readers' tastes change over time.

[69] Kathryn Zickuhr & Lee Rainie, *A Snapshot of Reading in America in 2013*, PEW RES. CTR. (Jan. 16, 2014), http://www.pewinternet.org/2014/01/16/a-snapshot-of-reading-in-america-in-2013/ (reporting that American adults read or listened to an average of 12 books in 2013).

half of readers, hitting a proverbial brick wall for the 25% of Americans who read no books at all. Here, traditional lowest common denominator strategies based on best-sellers will continue to dominate. This practically guarantees an ongoing role for intuition and human experts, although Big Data could still add significant insights.[70]

III. TRADITIONAL INSTITUTIONS: GETTING THE MOST FROM HUMAN JUDGMENT

Traditional publishing institutions evolved to fund and collect human judgment. Indeed, publishers brought new players (*e.g.* editors, bookstore owners) into the system if and only if their insights generated enough new sales to cover their costs. At the same time, evolution also led to multiple compromises, including high book prices and a profusion of middlemen. These erased many of the gains that improved predictions might otherwise have delivered to authors and readers.

A. *Ancient and Medieval Publishing*

The first written books date from the Fifth Century BC.[71] By the Second Century BC, there were so many titles that no single human could read them all.[72] Managing this effort led to three basic institutional solutions, all of which persist today.

Crowd-Sourcing. The simplest way to find out what readers enjoy is to ask them. Bards knew that an improvement that audience members suggested one night was worth trying again and, if that worked, the night after that. Given enough time, these iterative improvements could produce masterpieces like *The Odyssey*. Nor is this surprising: As computer programmers like to say, "given enough eyeballs, all bugs are shallow."[73] At the same time, relying on disorganized groups invited various pathologies:

> *Total Effort.* Improvements were limited to the small, "granular"[74] efforts that audiences were willing to donate in the course of a few hours around a campfire. This very slow process became intolerable once authors began writing books for career reasons like attracting

[70] Estimating the reactions of 1,000 reader groups is much easier than assigning 300 million Americans to specific groups. This is particularly true since reader groups sometimes share tastes, so that obtaining reactions from one group reliably predicts reactions from another group. *Cf.* SMITH & TELANG, *supra* note 30, at 335–36 (arguing that Big Data can identify potential blockbusters that traditional "gut feel" editors have overlooked).

[71] LIONEL CASSON, LIBRARIES IN THE ANCIENT WORLD 26 (2001).

[72] The last person who claimed to have read every book in existence died in the 2d Century BC. *See id.* at 38.

[73] *See, e.g.*, *Linus' Law*, WIKIPEDIA, https://en.wikipedia.org/wiki/Linus%27s_Law (last visited Apr. 12, 2018).

[74] See Benkler, supra note 51.

> students so that titles had to be completed ("published") in much less than a human lifetime.
>
> *Coordinating Effort.* Letting audience members independently decide which books to improve rewarded popularity. This implied a vicious cycle in which obscure texts were systematically neglected so that even fewer people read and improved them.
>
> *Merging Effort.* Then as now, many audience suggestions would have been worthless or mutually incompatible. This required crisp, unambiguous choices that no crowd could supply. Instead, the Homeric system delegated this final decision to bards, in much the same way that modern open source authorizes a kernel of distinguished developers to decide which code is accepted.[75]

Despite these defects, crowd-sourcing remains central to modern publishing. While we have emphasized that publishers often intervene to promote promising texts, they usually stop after the first few generations of readers. After that, the crowd has the last word.[76]

Literary Salons. The rise of celebrity authors required new institutions that could polish books in years instead of centuries. The main Roman response was to preview and revise manuscripts in elite salons.[77] This traded large audiences for intense effort by hard core enthusiasts. But salons still relied on volunteers.[78] This limited the supply of effort to what members were willing to donate or else could extract in more material benefits like enjoyment, prestige, or social contacts. Contemporary testimony suggests that the system was particularly bad at discovering new authors.[79] Lacking central direction, volunteers would have sampled some books many times while ignoring others completely.[80] Worse, volunteers who tried unknown titles had less time to enjoy those that had already been discovered and recommended. On the usual free-rider logic, this would have

---

[75] *See, e.g.*, Stephen M. Maurer & Suzanne Scotchmer, *Open Source Software: The New Intellectual Property Paradigm*, *in* 1 ECONOMICS & INFORMATION SYSTEMS 305 (2006).

[76] *See* Paul Elie, *supra* note 16.

[77] For a description of an elite salon, *see* WILLIAM JOHNSON, READERS AND READING CULTURE IN THE HIGH ROMAN EMPIRE: A STUDY OF ELITE COMMUNITIES 47 (2010).

[78] For a detailed description of the members who participated in a particularly famous salon headed by Fronto (AD 95 – 167), see *id.* at 137–56.

[79] Martial mocks young provincials who seek to "push [their] way among the great" as "mad," emphasizing that most are "pale with hunger." EPIGRAM, MARTIAL, BOOK III (1897); *see also*, GEORGE HAVEN PUTNAM, AUTHORS AND THEIR PUBLIC IN ANCIENT TIMES: A SKETCH OF LITERARY CONDITIONS AND OF THE RELATIONS WITH THE PUBLIC OF LITERARY PRODUCERS, FROM THE EARLIEST TIMES TO THE INVENTION OF PRINTING 250 (1893) ("[Martial] refers more than once to many amiable and deserving authors, who, despite their talents, succeeded in reaching no public at all . . . .").

[80] *Cf.* Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 J. L. & ECON. 265, 265–71 (1977) (introducing "prospect" model that justifies patents as a vehicle for coordinating R&D across multiple inventors).

encouraged them to skip search and let somebody else do the work.[81]

Like crowd-sourcing, salon methods still exist today, most notably in the way most authors seek out advice from readers.[82] If anything, digital technologies have expanded the practice by making casual on-line exchanges easier.[83]

Commercial Publishers. The defects of volunteer publishing cried out for more effort, hierarchy, and coordination. This made commercial publishing and paid employees a natural solution. Greek businessmen probably began making and selling books by the Fourth Century BC.[84] At first, however, merchants waited for someone to request a particular title. This changed in the high Roman period, when commercial publishers discovered that batch production was cheaper.[85] But producing books ahead of demand was a double-edged sword: Publishers whose titles failed to catch on could easily go broke. The archeological evidence suggests that this made publishers conservative, limiting the search for new titles and impoverishing readers.[86] Strangely, the system also produced a kind of informal copyright. Having a large inventory of copies on hand deterred would-be competitors from making and selling their own copies. This let publishers charge above-cost prices.[87]

The first publishers probably sold books through their own stores.[88] But publishers who found better ways to predict sales could

---

[81] This calculation was partly offset by the glamor of discovering forgotten texts. *See* Sarah Power, *The Curse of the Forgotten Authors*, THE GUARDIAN (Apr. 19, 2013), https://www.theguardian.com/books/booksblog/2013/apr/19/bestbookshops (a modern example).

[82] *See, e.g.*, ANDREW LYCETT, IAN FLEMING: THE MAN BEHIND JAMES BOND 298–99 (1995) (describing how Ian Fleming changed James Bond's pistol after a fan complained of 007's "'rather deplorable taste in firearms.'"). The academic system of previewing papers in seminars is an even more insistent echo of the Roman salons.

[83] *See e.g.*, Melissa Pearl, *Author Fan Club Awesomeness*, INDIES UNLIMITED (Mar. 27, 2015), http://www.indiesunlimited.com/2015/03/27/author-fan-club-awesomeness/.

[84] CASSON, *supra* note 71, at 27.

[85] *See e.g.*, Jon W. Iddeng, *Publica aut Peri! The Releasing and Distribution of Roman Books*, 81 SYMBOLAE OSLOENSES: NOR. J. GREEK & LATIN STUDIES 58, 63–64 (2006). There is good internal evidence that some, though not all, Roman manuscripts were produced in bulk, with a single reader dictating to a roomful of scribes. *Id.* at 65. This suggests that dictation could sometimes achieve lower unit-costs than sight-copying. On the other hand, publishers who used the method had to hire an additional worker. This implies a minimum efficient scale below which sight-copying continued to dominate.

[86] *Id.* at 64.

[87] To see why, put yourself in the place of a would-be copyist. Assume further that you know three things: (a) the title will sell exactly 100 copies, (b) the publisher has made 100 copies already, and (c) it is better to sell at a loss than to earn no money at all. It follows that if you make a 101st copy, the publisher will respond by selling his own copy below cost. But if you believe this, you will never enter the market in the first place. *See* Stephen M. Maurer, *From Bards to Search Engines: Finding What Readers Want from Ancient Times to the World Wide Web*, 66 S. CAROLINA L. REV. 495, 510–11 (2014).

[88] Casson notes that the first references to bookstores date from the late 5th Century BC, but that "we can only guess" whether they were independent from the "scriptoria" that produced books for sale. *See* CASSON, *supra* note 71.

earn more. This meant, among other things, paying outsiders to contribute their information and insights. The first such players were independent bookstores that earned a living by buying and reselling books at a markup.[89] The use of arm's length transactions had two important advantages. First, bookstores only made a profit if they found customers. This meant that any copies they bought from publishers represented their best estimate of what customers would actually buy. Second, readers knew that publishers had every reason to hype their own titles regardless of quality.[90] The existence of independent bookstores that carried the best books regardless of publisher eliminated this risk and boosted sales.

Independent bookstores also changed the structure of publishing. On the familiar logic of trademark, customer trust gave stores market power.[91] This, however, let them raise prices and may have driven some consumers out of the market. The saving grace, in this pre-industrial era, was that books were often sold through haggling.[92] This gave store owners an incentive to give less enthusiastic readers price breaks so that they bought books after all.

B. *The Age of Print*

The ancient ecosystem collapsed with the Fall of Rome. After that, mass publishing ceased to exist until the rise of universities in the 11th Century.[93] This touched off three distinct revolutions that shaped publishing throughout the age of paper.

The first development was the rediscovery of bulk production and scale economies. This led to steadily falling prices that expanded sales beyond universities to the general public.[94] The invention of print technology in the 15th Century accelerated this trend. Meanwhile, bulk production also resurrected the old Roman tactics for deterring copyists. The resulting above-cost pricing meant that editors who searched out and improved texts could recover their costs by tithing readers for each

[89] Iddeng, *supra* note 85, at 77 ("A reasonably sized *civitas* might have a cornerstore that sold some books as well (such as Lyon had), and now and again a travelling salesman brought fresh books to town.").

[90] One silver lining is that readers had no reason to mistrust each publisher's *relative* ranking of books *within* its own catalog. *See* Fadiman, *supra* note 23.

[91] This was over and above the market power that stores enjoyed from simple geographic scarcity.

[92] The detailed history is complex. While there is very little evidence of pricing in ancient and medieval times, books sellers from the 17th Century onward frequently tried to cartelize prices. These attempts were undercut by sellers like James Lackington (1746-1815) who refused to respect publishers' announced resale prices. *See generally*, *Books as a Commodity*, THE BOOK: 1450 TO THE PRESENT, http://eduscapes.com/bookhistory/commodity/index.htm (last visited Apr. 25, 2018).

[93] PUTNAM, *supra* note 79, at 66, 238, and 291.

[94] *Id.*

copy sold.[95]

The second revolution expanded the publishing ecosystem far beyond Roman precedents. This was particularly urgent in an era when the high fixed costs of bulk production – and later print – punished publishers who picked the wrong titles. Probably the most novel institution was the Frankfurt book fair, where merchants met each year to trade inventory which they then sold across Europe.[96] Publishers also brought human judgment in-house, hiring specialist "literary counsellors" and "triers" from the Eighteenth Century onward.[97] For their part, bookstores began purchasing book reviews from professional critics. By the late 20th Century, the system was generating about 50,000 reviews per year.[98]

The final innovation, following the Statute of Anne (1710)[99], was the introduction of formal copyright statutes. Somewhat paradoxically, this legal innovation was similarly anchored in the physical costs of print which (a) limited the number of pirates, and (b) guaranteed that publishers would have assets to seize if they obtained a judgment.[100] Copyright, in turn, made new business models possible. Publishers could now offer whole portfolios of titles in small first editions and then wait to see which ones caught on. This led to a wild proliferation of titles from the Enlightenment onward.[101]

C. *Modernity*

Book consumption exploded from the start of the Nineteenth Century.[102] This drastically complicated search, forcing publishers to reach beyond elite tastes to find titles that would please wider audiences including tradesmen and servants.

Market Power. The large fixed costs associated with print and

---

[95] *Aldus Manutius*, WIKIPEDIA, https://en.wikipedia.org/wiki/Aldus_Manutius (last visited Apr. 12, 2018).

[96] NICOLE HOWARD, THE BOOK: THE LIFE STORY OF A TECHNOLOGY 76 (2005).

[97] SHER, *supra* note 44, at 283.

[98] Joel Waldfogel, *Copyright and Technological Change in Music, Movies, and Books, in* 2 RESEARCH HANDBOOK ON THE LAW & ECONOMICS OF INTELLECTUAL PROPERTY (Peter S. Menell, David L. Schwartz & Ben Depoorter, eds., forthcoming 2018).

[99] Act for the Encouragement of Learning (Statute of Anne), 8 Ann. c. 21. (1710) (Gr. Brit.).

[100] The Romans understood the concept of copyright but never implemented it. *See, e.g.*, PUTNAM, *supra* note 78, at 268. This was presumably because such a law would have been unenforceable in a world where most books were still made by amateur copyists. This logic was repeated in our own time by Soviet authorities' failure to suppress "Samizdat" manuscripts despite draconian restrictions on access to Xerox machines and even typewriters. This shows that even police states find it hard to suppress non-commercial publishers. *See generally*, VICTOR SEBESTYAN, REVOLUTION 1989: THE FALL OF THE SOVIET EMPIRE 164 (2009) (describing Rumanian regulation of typewriters).

[101] SHER, *supra* note 44, at 2.

[102] George Urwin, Philip Soundy Unwin & David H. Tucker, *History of publishing*, ENCYCLOPAEDIA BRITANNICA (last updated Mar. 3, 2018), https://www.britannica.com/topic/publishing#ref28633.

marketing made concentration inevitable. But there were also dynamic forces at work. Readers who received good titles from publishers were more likely to buy a second time. This, however, favored big publishers with large catalogs. The effect was particularly strong for bestsellers, which were already dominated by today's household names at the start of the 20th Century.[103] One hundred years later, the "Big Five"[104] have largely suppressed price competition among themselves,[105] although they probably compete on search.[106] They also use their extensive copyright portfolios to suppress older titles going back to the Twenties, which are now much less available in digital form than their Victorian predecessors.[107]

Nobody should be surprised if these giants changed the rules to suit themselves. This included imposing standard resale prices that largely eliminated traditional haggling between booksellers and readers. By the late Thirties, the Big Five launched a new price discrimination scheme that sold identical texts as expensive "hardbacks" alongside much cheaper, "mass market" paperback editions.[108] Naively, one might have guessed that cheap paperbacks would destroy the hardback market. But in fact, consumers showed a fierce preference for hardbacks for titles they valued or hoped to reread. Like haggling, this strategy had the virtue of simultaneously expanding profits and readership. Indeed, the first paperback mysteries sometimes sold nearly ten times what hardbacks did.[109]

---

[103] For example, all but two of the top ten bestsellers for 1900 were published by ancestors of today's Big Five publishers. *Cf. Publishers Weekly List of Bestselling Novels in the United States in the 1900s*, WIKIPEDIA, https://en.wikipedia.org/wiki/Publishers_Weekly_list_of_bestselling_novels_in_the_United_States_in_the_1900s#1900 (last visited Apr. 12, 2018); *see also Search*, OCLC WORLDCAT DATABASE, https://www.worldcat.org/account/?page=searchItems (last visited Apr. 12, 2018).

[104] Today's "Big Five" were called the "Big Six" prior to the merger of Penguin and Random House in 2013.They compete with an estimated 3-400 smaller publishers worldwide. Joost Poort & Nico van Eijk, *Digital Fixation: The Law and Economics of a Fixed E-Book Price*, 23 INT'L J. CULTURAL POL'Y 464 (2015).

[105] By the early 21st Century, this advantage was so large that the Big Five charged twice as much for e-books as other publishers. *The B&N Report*, AUTHOR EARNINGS, http://authorearnings.com/report/the-bn-report/ (last visited Apr. 12, 2018).

[106] U.S. v. Apple, Inc., 791 F.3d 290, 300 (2d Cir. 2014) (reporting that the CEOs of the Big Five " . . . 'did not compete with each other on price,' but over authors and agents").

[107] *See, e.g.*, Paul J. Heald, *The Demand for Out-of-Print Works and Their (Un)Availability in Alternative Markets*, ILL. PUB. L. & LEGAL THEORY, 17–19 (2014); *see also* Imke Reimers, *Copyright and Generic Entry in Book Publishing* (Nat'l Bureau of Econ. Research, Working Paper, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2938072 (copyrighted works cost 35% more on average compared to public domain titles).

[108] Ann Trubek, *How the Paperback Novel Changed Popular Literature*, SMITHSONIAN (Mar. 30, 2010), https://www.smithsonianmag.com/arts-culture/how-the-paperback-novel-changed-popular-literature-11893941/.

[109] Raymond Chandler's hardback edition of *The Big Sleep* sold 4,000 copies in its first nine months; the paperback sold 300,000. TOM WILLIAMS, A MYSTERIOUS SOMETHING IN THE LIGHT: THE LIFE OF RAYMOND CHANDLER 175, 178 (2012). The paperback revolution also created space for new authors and niche titles that publishers never issued in hardback at all. *See*

Ecosystem. Modern print publishers lived or died by their ability to predict which titles would sell a thousand or so copies.[110] Starting in Victorian times, this led to a steady expansion of booksellers, libraries, newspaper reviewers, book-of-the-month clubs, literary agents,[111] and other middlemen. The market power of these new players sometimes approached that of publishers themselves.[112]

But book stores also had to predict which books to buy, and how long to let them gather dust on costly shelf space. The problem, as market power shifted back to publishers at the start of the 20th Century, was that bookstores had less and less margin for mistakes. This reached crisis proportions in the 1930s, when publishers intervened to reduce stores' downside risk by promising to buy back unsold copies.[113] The system was and remains astonishingly costly: Fully forty percent of today's physical books are eventually discarded and pulped.[114]

The new regime lasted into the 1980s, when the rise of Big Box bookstores began shifting market power back to retailers.[115] Soon, these new players were forcing down hardcover prices for bestsellers. This enticed foot traffic into the stores in much the same way that cheap digital titles would later sell e-readers. But it also eliminated publishers' ability to price discriminate. The result, in the 21st Century, is that traditional "mass market" paper editions have mostly been replaced by higher-priced "trade paper" copies that many readers refuse to buy.

Should I be Worried if my Book is Published as a Paperback Original?, THE WRITER (Apr. 28, 2017), https://www.writermag.com/2017/04/28/paperback-original/.

110 Piersanti, *supra* note 3 ("average book generates $50,000 to $150,000 in sales" and sells fewer than 2,000 copies over its lifetime).

111 Most large publishers only accept manuscript submissions through "established" literary agents. *See, e.g., Frequently Asked Questions*, HACHETTE BOOK GROUP, https://www.hachettebookgroup.com/about/faqs/#submissions (last visited Apr. 25, 2018). The practice amounts to outsourcing significant editing tasks, presumably including a markup reflecting the scarcity value of "established" agents.

112 Maurer, *supra* note 14, at 523 (dominant role of British commercial libraries in British publishing from mid-19th to mid-20th Centuries) and 544 (market power of "big box" stores from the 1990s onward); Greg Ip, *The Antitrust Case Against Facebook, Google and Amazon* WALL STREET J. (Jan. 17, 2018), https://www.wsj.com/articles/the-antitrust-case-against-facebook-google-amazon-and-apple-1516121561 (Amazon site accounts for roughly 75% of all ebook sales).

113 Lyyn Neary, *Transcript: Publishers Push for New Rules on Unsold Books*, NPR (June 13, 2008), https://www.npr.org/templates/story/story.php?storyId=91461568 ("During the Great Depression, publishers were looking for a way to encourage booksellers to buy more books and to take a chance on unknown authors. So they offered bookstores the right to return unsold books for credit.").

114 George P. Landow, *Review of Griest's Mudie's Circulating Library and the Victorian Novel*, 69 MODERN PHILOLOGY 367–69 (1972); GUINEVERE L. GRIEST, MUDIE'S CIRCULATING LIBRARY AND THE VICTORIAN NOVEL (1970); *Charles Edward Mudie*, WIKIPEDIA, http://en.wikipedia.org/wiki/Charles_Edward_Mudie (last visited Apr. 4, 2018).

115 Maurer, *supra* note 14, 544–45.

D. *Welfare*

We have seen that commercial publishing featured monopoly pricing and a proliferation of middlemen. This led to several benefits, the usual IP defects, and multiple distortions.

Benefits. Commercial institutions improved on older, volunteer-based search for two basic reasons. First, paid employees were far more willing to accept unpleasant tasks and do what they were told. This allowed a massive increase in specialization, hierarchy, and coordination which, in turn, opened the door to new publishing models that featured intense effort by a handful of experts. Second, publishers only needed to discover a winning title once. After that, they could tithe each reader who bought the book to recover their search costs. These two circumstances – that search only needs to be performed once, and can be paid for by spreading small tithes across large audiences – became the model for IP business models ever afterward.

Monopoly Distortion. The conventional objection to IP is that it supports high prices leading to lost sales and less consumption. However, the special circumstances of cultural markets also produced a more subtle pathology. By almost any measure, the number of people who read a book in any one year is much less important than total readership over the centuries.[116] This implies that an ideal publishing system should respect quality rather than age so that, in Hemingway's phrase, "the good writer competes only with the dead."[117] But copyright fails this test. While public titles written before 1923 are routinely available as e-books, most copyrighted midcentury works are not.[118] The reason almost certainly is that new digital titles command five-to-six dollar markups compared to just one or two dollars for older works.[119] This encourages the Big Five to suppress older titles that might hurt ("cannibalize") the market for new offerings.[120]

Proliferation of Middlemen. The traditional publishing industry's defects do not end with IP. Instead, they have been greatly amplified by the dispersion of market power across multiple independent actors.

116 There is also the important externality that longevity creates culture-defining "classics" that provide a common socialization for readers.

117 *Quoted in* RAYMOND CHANDLER, THE SIMPLE ART OF MURDER 53 (2002).

118 The existing deficit also implies that there is little incentive for editors to invest the time and effort that led earlier generations to rediscover, for example, Melville and James.

119 *May 2015 Author Earnings Report*, AUTHOR EARNINGS, http://authorearnings.com/report/may-2015-author-earnings-report/ (last visited Apr. 12, 2018).

120 Maurer, *supra* note 14; Erik Maskin & John Riley, *Monopoly with Incomplete Information*, 15 RAND J. OF ECON.171,175, 189–90 (1984); Michael Mussa & Sherwin Rosen, *Monopoly and Product Quality*, 18 J. OF ECON. THEORY 301, 304–307 (1978); *see also*, U.S. v. Apple, Inc., 791 F.3d 290, 342 (2d Cir. 2014) (Judge Jacobs dissenting, arguing that low-priced e-books "cannibalized" sales of more profitable hardcover editions). The logic of suppression is even stronger to the extent that Big Five members behave as a cartel, so that each member acts *both* to avoid cannibalizing its own titles *and* those of its competitors.

Economic theory emphasizes that goods that reach consumers after multiple markups will be priced above the price that a single monopolist would charge.[121] This overpricing hurts readers, but also authors, publishers, bookstores and every other actor in the system.

The proliferation of actors also creates incentives to deceive, most notably by encouraging publishers to exaggerate the quality of their own titles and even come out with second-best "me too" titles that they know to be inferior. One partial fix is for publishers to buy back physical copies that fail to catch on.[122] But this is itself wasteful. Meanwhile, digital markets have created still more middlemen by subdividing retailers into both physical and on-line stores. This has led, among other things, to a "showrooming" externality in which consumers examine titles in physical bookstores but order them on-line.[123] As a result, the costlier but more informative physical channel has become increasingly unsustainable.[124]

Table 1: Commercial Publishing Pathologies

Problem 1: Monopoly	
	High prices Leading to Lost Sales and Readership
	Suppression of Older Works
	Wasteful enforcement methods leading to "pulped" and discarded books.

[121] AUGUSTIN COURNOT, RESEARCHES INTO THE MATHEMATICAL PRINCIPLES OF THE THEORY OF WEALTH (Nathanial T. Bacon trans., 1838). To see why, imagine a shoe monopoly. In the benchmark case of a single monopolist, we expect the seller to continue raising shoe prices to the point where further increases lose more revenue from driving consumers out of the market than they gain in extracting higher payments from those who remain. Now consider what happens when the market is divided between independent left- and right-shoe monopolists. When the left-shoe monopolist raises prices, she only feels half the lost revenue: The rest falls onto the right-shoe monopolist. But she still receives the full markup from each transaction. This leads to a Prisoner's Dilemma result in which both sides raise prices far above the single monopolist price. This impoverishes not just consumers but also the monopolists themselves. In this Alice-in-Wonderland world, cartelization produces lower prices and more welfare.

[122] Publishers also need to have paper copies on hand so that readership can quickly expand if and when a title starts to "take off." Consumers who are told that a particular title is out of stock seldom have much trouble obtaining their second-best choice elsewhere. This probably explains why e-books sell 40% fewer copies when publishers delay publication until months after the hardback release. SMITH & TELANG, *supra* note 30, at 95–96.

[123] *Showrooming*, WIKIPEDIA, https://en.wikipedia.org/wiki/Showrooming (last visited Apr. 25, 2018).

[124] Piersanti, *supra* note 3 (noting "the disappearance over the past decade of over 500 independent bookstores and the Borders bookstore chain…").

Problem 2: Middlemen	
	Double Monopoly Problem
	Publication of second-best, "Me-Too" Titles
	Deliberate Overproduction and Pulping of Books
	"Showrooming" Externalities
	Divided control makes price discrimination difficult or impossible.

IV. FIRST FUTURE: INCREMENTALISM

Publishers have tried many different institutions over the past 2,000 years. At this late date, we should be skeptical that anyone will find significantly better ways to harvest human judgment. The difference for Big Data is that it promises to make expert judgments and the institutions that fund them superfluous. The next three sections ask what is likely to replace them.

A. *Post-Modern Publishing*

The simplest technological forecast is stasis, *i.e.* that Big Data methods will never get much better than they are today. If so, human editors will remain the gold standard for predictions, even if Big Data's "readers who like X also liked Y" are sometimes more cost-effective. This is more or less the world we live in, though the revolution is still not complete.

Diminished Copyright. Commercial publishers historically relied on large up-front costs to deter pirates, both directly in the market place and indirectly, by facilitating copyright. For the past quarter century, however, these physical costs have steadily eroded in the face of digital book production (1980s), print-on-demand and on-line bookstores (1990s), and e-books (2000s). Remarkably, up-front costs are now so low that college students can commit large-scale piracy from their dorm rooms.[125]

And yet, remarkably, the sky has not fallen. Instead, the logic of protection has migrated to transactions costs.[126] Provided that legitimate

[125] U.S. v. LaMacchia, 871 F.Supp. 535 (D.Mass. 1994).

[126] For a dramatic example, see Wayne Ma, *How a Plague on the Movie and Music Industries Became Their Chief Protector in China*, WALL STREET J. (May 21, 2017),

titles are not priced "too high," readers still prefer to avoid the time and effort that would be needed to search out illicit copies.[127] In this new system, copyright is mainly needed to block pirate sites from becoming so open and notorious that their convenience would match legitimate retailers. This mechanism is decidedly weaker than the old copyright, but neither is it zero.

The New Search. The new digital technologies did more than enable pirates. They also slashed manufacturing and inventory costs. The result is that print-on-demand and e-book publishers only need one or two sales to break even, while on-line stores can stock hundreds of thousands of titles at essentially no inventory cost. This has drastically reduced the downside risk of picking poor titles, even if the jackpot from discovering the next bestseller remains as attractive as ever. At least potentially, the upside benefits are even more important. We have argued that modern copyright encouraged publishers to issue large numbers of titles in small initial print runs hoping that one or two would catch on. This massively contributed to The Enlightenment's explosion of titles. Digital technologies, which drive fixed costs nearly to zero, have made such portfolio strategies even more attractive, always supposing that piracy threats can be contained.

The downside, for publishers, is that digital technologies erode copyright. All else equal, we should therefore expect less search and marketing effort. There is strong, if anecdotal evidence that this is happening. First and most strikingly, unknown authors now sell fewer e-books than hardcovers. Evidently, publishers have less budget to find and promote digital titles than they do for paper.[128] Second, today's publishers often withhold book contracts until *after* titles have been self-published and racked up sales.[129] This shows that commercial

https://www.wsj.com/articles/how-a-plague-on-the-movie-and-music-industries-became-their-chief-protector-in-china-1495364406 (quoting recording industry official Neil Turkewitz, "Baidu almost single-handedly eroded the value of music [in China]").

127 Maurer, *supra* note 87; *see also*, SMITH & TELANG, *supra* note 30, at 209–10, 215 (arguing that consumers prefer convenience and discussing the piracy tradeoff). The average US hourly wage is approximately $26. *See Economic News Release: Table B-3. Average Hourly and Weekly Earnings of All Employees on Private Nonfarm Payrolls by Industry Sector, Seasonally Adjusted*, BUREAU OF LABOR STATISTICS, https://www.bls.gov/news.release/empsit.t19.htm (last visited Apr. 12, 2018). This makes it cheaper for most readers to purchase a legitimate book for $10 than to spend 24 minutes looking for copy that costs nothing at all.

128 Jeffrey A. Trachtenberg, *Authors Feel Pinch in Age of E-Books*, WALL STREET J. (Sept. 26, 2010), https://www.wsj.com/articles/SB10001424052748703369704575461542987870022; *see also*, Ben Fritz, *For Movie Producers, A Golden Age Fades: As Hollywood Slashes Spending, Nobody Has Felt the Burn as Much as Movie Producers*, WALL STREET J. (Jan. 22, 2014), https://www.wsj.com/articles/for-movie-producers-a-golden-age-fades-1390016141 (noting that studios have fewer search resources).

129 The first example was *Fifty Shades of Grey* in 2011. Since then, about ten percent of all top-ten bestsellers have begun life as self-published books. Joel Waldfogel & Imke Reimers, *Storming the Gatekeepers: Digital Disintermediation in the Market for Books*, 31 INFORMATIONAL ECON. AND POL'Y 56 (2015). It is worth asking why an author whose books

publishers do not even try to screen three-quarters of the titles that appear each year. Finally, modern bestseller lists contain repeat authors sixty percent more often than they did in the 1960s.[130] This implies that readers are using proxies like past success more often – and that new authors find it harder than ever to get noticed.

So far, we have argued that reduced corporate budgets have impoverished search. But, on closer examination, publishers have not so much abandoned the task as offloaded it onto authors[131] and readers. The model has been greatly facilitated by Big Data tools that replicate some of the simpler logic that bookstore clerks used to provide. While this sacrifices their shrewder insights, Amazon's twenty-fold improvement in inventory has ensured that high-end consumers barely notice the loss.[132]

The question remains what to do about the many less enthusiastic readers for whom search is not worth the effort. Probably the most revolutionary feature of Amazon's new brick-and-mortar stores is that they shelve books facing outward.[133] The waste of space only makes sense if consumers prefer Amazon recommendations to scouring traditional stores that stock more titles. This is essentially the old bestseller pattern except that Amazon picks have replaced the Big Five. But that could still be an improvement compared to traditional bestsellers, in the same way that Netflix now finds it profitable to offer niche content that traditional TV hardly ever broadcast.[134]

---

have begun break through would even want a publisher. Presumably they worry that fickle markets might still drop them later on.

130 Returning authors constituted about fifty percent of top ten weekly bestsellers in 1961−1970, but 80% from 2007−2016. *See Publishers Weekly List of Bestselling Novels in the United States*, WIKIPEDIA,https://en.wikipedia.org/wiki/Publishers_Weekly_lists_of_bestselling_novels_in_the_United_States (last visited Apr. 12, 2018).

131 Piersanti, *supra* note 3.

> Publishers have managed to stay afloat in this worsening marketplace only by shifting more and more marketing responsibility to authors, to cut costs and prop up sales. In recognition of this reality, most book proposals from experienced authors now have an extensive (usually many pages) section on the authors' marketing platform and what the authors will do to publicize and market the books. Publishers still fulfill important roles in helping craft books to succeed and making books available in sales channels, but whether the books move in those channels depends primarily on the authors.

Id.

132 Erik Brynjolfsson et al., *supra* note 7 (estimating that US consumers gained $1bn in annual consumer surplus from on-line stores' radically expanded title lists); JOEL WALDVOGEL, THE RANDOM LONG TAIL AND THE GOLDEN AGE OF TELEVISION, in INNOVATION POLICY AND THE ECONOMY 25 (Josh Lerner & Scott Stern eds., 2018) (increased consumer choice has more than made up for lost guidance).

133 Alexander Alter & Nick Wingfield, *A Trip Through Amazon's First Physical Store*, N.Y. TIMES (Mar. 10, 2016), https://www.nytimes.com/2016/03/12/business/media/a-virtual-trip-through-amazons-physical-store.html.

134 Adam Levine-Weinberg, *How Netflix Really Creates Value*, MOTLEY FOOL (Sept. 30, 2015), https://www.fool.com/investing/general/2015/09/30/how-netflix-inc-really-creates-value.aspx ("For the right price, Netflix can afford to buy content that doesn't have broad popularity, because it can target particular subscribers who may be interested via personalized

Subscriptions and Search. We have said that independent bookstores improved search by giving readers a trusted guide to sorting out publishers' hype. Digitization lets stores take this one step further by making all titles available for a single subscription.[135] This all-you-can-eat model has already been implemented by several large book services including Kindle Unlimited,[136] Oyster,[137] and Scribd.[138] The scheme offers at least three advantages. First, subscriptions (unlike independent bookstores) do more than just protect readers against dishonesty. They also protect against honest (but mistaken) recommendations since it now costs nothing for consumers to discard bad titles and try again.[139] Second, increased sampling by consumers means more shopping and more data for Big Data to feed on. This is one of the cheapest ways for companies to expand the supply of objective information. Finally, subscriptions mix old and new titles indiscriminately.[140] This automatically implements Hemingway's advice that all titles should compete on an even footing.

The downside, of course, is that readers will normally prefer whichever service offers the most titles. This invites the usual rich-get-richer dynamic that retailers succeed *because* they are large and become monopolists even faster. Worse, readers who have already bought subscriptions are less likely to seek books elsewhere. This gives dominant retailers yet another lever for punishing publishers that defy them.[141]

Room for Improvement. Despite these developments, the revolution is not complete. Probably the biggest surprise is that e-books have not replaced paper. Instead, both technologies seem stable and

recommendations. 'Linear' TV networks can't even consider broadcasting niche content except at very odd hours.").

135 Like most publishing models, there are important antecedents, notably including the large commercial libraries that flourished in Britain from Victorian times until World War II. Maurer, *supra* note 87, at 523.

136 *Kindle Unlimited*, AMAZON, https://www.amazon.com/gp/feature.html?ie=UTF8&docId=1002872331 (last visited Apr. 12, 2018).

137 Google-owned Oyster offers over 1,000,000 titles from over 1,600 publishers, including all of the Big Five. *Oyster (company)*, WIKIPEDIA, https://en.wikipedia.org/wiki/Oyster_ (last visited Apr. 12, 2018) (company).

138 *About Us*, SCRIBD, https://www.scribd.com/about (last visited Apr. 12, 2018) (advertising millions of books).

139 SMITH & TELANG *supra* note 30, at 104 (arguing that subscription cable allowed consumers to discover movies "they weren't willing to pay $15 to see in theaters.").

140 *See, e.g.*, *Kindle Unlimited*, *supra* note 136 ("Around 500 public domain titles are included in Kindle Unlimited, all of which we've synched with their free audiobook companions as a benefit to Kindle Unlimited subscribers.").

141 Amazon has been widely accused of denying its market to recalcitrant publishers. Evan Hughes, *Bringing Down the Hachette*, SLATE (May 30, 2013), http://www.slate.com/articles/technology/technology/2014/05/amazon_hachette_dispute_how_the_big_five_publishers_could_have_avoided_the.html.

poised to coexist for years.[142] Moreover, the fact that e-readers have colonized essentially the same throwaway genres (*e.g.* mysteries, romance) that cheap paperbacks used to serve implies that both systems are grounded in some deep-seated and durable split in why readers buy books in the first place. But in that case, we should expect the new technology to reinstate something like the old mid-century price discrimination.[143] The practical obstacle, compared to fifty years ago, is that the Big Five now control high-end (hardback) pricing but have ceded low-end (e-book) control to Amazon. The standoff has only hardened since 2013, when the US Justice Department intervened to stop the Big Five from taking back e-book pricing from Amazon.[144]

But even if price discrimination is restored, it is unlikely to last. In the long run, a really convincing version of electronic paper will give readers devices that (a) feature the look-and-feel of books, but (b) can be endlessly reprogrammed with new content. At that point, hardbacks will disappear entirely, so that any remaining differences between different editions will be entirely digital. This will leave much less room for today's "second degree" price discrimination strategies based on versioned goods.[145] Conversely, it will encourage publishers to revisit so-called "first degree" strategies based on estimating readers' willingness to pay directly, though it is not at all clear that today's primitive Big Data methods would be up to the task.

B. *Welfare*

The digital revolution let publishers offload search onto consumers at the same moment that piracy was forcing down book prices. This coincidence turned out to be a good trade for consumers and social welfare. The question is whether policymakers can go beyond this happy accident to engineer even more favorable exchanges.

Adjusting Copyright. Given the recent erosion of copyright, it is natural to think that reforms should push back by increasing reward or enforcement. But theory is indeterminate. Bigger copyright rewards

---

142 *See, e.g.*, Carolyn Kellogg, *6 Book Trends for 2016: Look into the Future*, L.A. TIMES (DEC. 31, 2015), http://www.latimes.com/books/la-ca-jc-book-trends-20160103-story.html ("Meanwhile, e-books, which were once predicted to reach 50% to 60% of total book sales, hovered at just 25%."); Piersanti, *supra* note 3 ("After skyrocketing from 2008 to 2012, e-book sales leveled off in 2013 and have fallen more than 10% since then).

143 The main difference compared to mid-century price discrimination model is that "windowing" strategies that suppress e-books for months after hardback releases no longer work. U.S. v. Apple, Inc., 791 F.3d 290, 301 (2d Cir. 2014); SMITH & TELANG, *supra* note 30, at 95–96 (delayed release suppressed book sales by 40 percent. Missing consumers either permanently lost interest or pursued pirated editions).

144 *See infra* Section X.A.

145 Second degree price discrimination confronts consumers with quality choices that have been deliberately engineered to make high-willingness-to-pay consumers reveal themselves. *Price Discrimination*, WIKIPEDIA, https://en.wikipedia.org/wiki/Price_discrimination (last visited Apr. 12, 2018).

might re-expand corporate search programs, delivering more and better choices to readers. But we can equally imagine the industry pocketing the rents and providing very little. What policymakers need now is empirical guidance. Probably the most promising strategy would be to conduct fiscal experiments that subsidize selected titles to see whether publishers really would market them harder.

The Unfinished Revolution: Price Discrimination. The rise of mass-market paper formats in the mid-20th Century dramatically expanded readership while protecting and enhancing IP royalties.[146] We have argued that e-books could restore a similar system today. The impediment, for now, is that the *Apple* decision has slammed the door on publishers' attempts to reassert control over e-book prices. We stress below that the doctrinal issue is subtle, and it is possible to imagine the Second Circuit deciding the same way. The problem so far is that neither the Second Circuit nor the US Justice Department shows any indication of having ever noticed the issue.[147] Whatever else one might think, price discrimination is so central to economic efficiency that it should not be decided by inattention.

V. SECOND FUTURE: REIMAGINING PROPRIETARY MODELS

Future advances in Big Data will trigger a radical simplification of today's ecosystem. The only question is which parties will survive. This Section analyzes the most likely scenario in which today's dominant on-line retailers use their advantages in wealth, numeracy, and market power to push out or marginalize everyone else.[148]

A. *Vertical Integration*

Amazon's recent expansions into upstream publishing[149] and downstream brick-and-mortar stores[150] show that vertical integration is already profitable. Further improvements to Big Data will only increase these incentives. There are essentially three reasons for this:

---

[146] *See* discussion *supra* Section III.C.

[147] This is evident from, among other things, my conversations with Berkeley colleagues close to the case.

[148] We assume for analytical convenience that Amazon's expansion will take a conventionally proprietary form. That said, there are good strategic reasons why the company might bankroll an *open* search platform instead. These include, among other things, (a) assuring customers that recommendations were honest, (b) arousing fewer antitrust suspicions than a conventionally proprietary search facility, and (c) harvesting large amounts of volunteer labor. The strategy is even more attractive since most of the profits from expanded readership would return to Amazon as a monopolist in any case. Lest this scenario seem fanciful, IBM has invested more than a billion dollars in the open source Eclipse collaboration for closely analogous reasons. *See* Stephen M. Maurer, *The Penguin and the Cartel: Rethinking Antitrust and Innovation Policy for the Age of Commercial Open Source*, 1 UTAH L. REV. 269, 271 (2012). We return to the possibility of open platforms in Section VII below.

[149] U.S. v. Apple, Inc., 791 F.3d 290, 343 (2d Cir. 2014).

[150] Alter & Wingfield, *supra* note 133.

> *Market Efficiency*. The current ecosystem depresses sales (a) by inserting multiple middlemen and markups, and (b) promoting pathological competition based on hype, me-too books, and similarly dishonest methods. Fixing these problems will automatically expand the market and generate more profits.
>
> *Scale Economies in Data*. Big Data favors large databases.[151] But in that case, firms with large customer bases can offer better predictions and attract still more customers until they become monopolists. Clever sharing arrangements between rivals can avoid this, but only if policymakers insist on architectures that avoid cartelization.[152]
>
> *IT Costs*. Big Data will require large information technology investments. Developing software within a single organization greatly simplifies matters. This gives monopolists an inherent cost advantage.

The question remains whether vertical integration will continue until just one firm stands between authors and readers. In practice, publishers may never disappear entirely. First, we have already argued that on-line retailers have very limited capacity to integrate Big Data and human insights in-house. This suggests that Coasian strategies based on arms-length transactions by independent firms could yield better predictions. Second, publishers do more than just search. They also help authors improve and, especially, shape their manuscripts to deliver what readers want. Big numerate firms will find it hard to automate or recreate these functions in-house. Finally, we have said that Big Data and human judgment sometimes develop different insights. The large premiums that today's readers already pay for academic books suggests that they may value these added insights enough to tolerate the cost of human editors.[153] At the other end of the scale, bestsellers present a similar situation in which even slightly better insights would be well-worth the cost.

B. *Simplifying the Ecosystem*

We have emphasized that middlemen are costly, and that clearing them away can improve economic efficiency. The rub is that this same integration will equally extend the dominance of today's on-line

---

151 *See supra* Section II.B.

152 *See infra* Section IX.B.

153 The rise of publicly-available scholarly depositories like ArchivX (physics) and Social Science Research Network (law and economics) has done almost nothing to displace traditional journals that depend on humans to identify and print the best "archival" research. One natural interpretation is that academic audiences need and value functions like gatekeeping (*i.e.* identifying the best contributions) and curation (adding relatively small corrections to texts). This says nothing, of course, about whether the current system could be organized more efficiently. *See, e.g.*, *The Serials Crisis*, WIKIPEDIA, https://en.wikipedia.org/wiki/Serials_crisis (last visited Apr. 12, 2018).

retailers to up- and downstream markets.

New Horizontal Monopolies. Vertical integration would replace up- and downstream-competition with a single dominant actor. But is this really a loss? On the upstream end, the current system pairs (imperfect) competition among publishers with Amazon's on-line distribution monopoly. But in that case Amazon should be able to rake off any savings before they ever reach consumers. If Amazon dislikes this arrangement, it must be because it thinks that it can do the job better and more cheaply than the publishers themselves.

The situation for brick-and-mortar bookstores is more nuanced. We would indeed prefer the old system of competitive booksellers. The problem, by all accounts, is that it is no longer sustainable. This gives Amazon very little choice about moving into physical bookstores – if it wants the showrooming, it will have to pay for it.

Price Discrimination Revisited. Vertical integration would immediately end the tug-of-war over e-book prices and give Amazon unchallenged power to practice price discrimination. But it would *also* change Amazon's focus by giving it a stake in the hardback market. At that point, the company would set e-book prices based on the *combined revenue* from e-readers *and* hardbacks. Depending on the relative size of these markets, we would expect the company to set prices that promote growth across both channels.

There is also a more dramatic possibility. In the long run, any Big Data system good enough to predict readers' tastes should equally be able to estimate the intensity of those tastes. This would let companies tailor prices to individuals for the first time since books were sold by haggling[154] so that the familiar "mass market" dissolved into millions of bilateral transactions in which each reader received completely different prices and recommendations from every other reader.[155]

Mobilizing Readers. Depending on how Big Data evolves, sellers could decide that they need more data than can be obtained *gratis* in the course of ordinary business operations. One obvious method is to pay readers to review selected titles.[156] But, in that case, readers might

---

154 Today's small e-publishers already do this by repeatedly posting the same texts at slightly different prices. The tactic may be based on a judgment that high willingness-to-pay customers will buy immediately, leaving less enthusiastic customers to search for cheaper editions. Alternatively, the very small differences could be an experimental attempt to probe demand elasticity. *See* Maurer, *supra* note 14.

155 Netflix argues that it uses Big Data less to pick winning content than to target those who are likely to enjoy it. SMITH & TELANG, *supra* note 30, at 337.

156 This is the same strategy that Xerox Parc used in the Eighties when it radically subsidized in-house IT costs so that employees would try out new ways to use computers. *See e.g.*, WARREN TEITELMAN, THE CEDAR PROGRAMMING ENVIRONMENT: A MIDTERM REPORT AND EXAMINATION, PALO ALTO RES. CTR. (June 1984) 1, 8, *found at* https://ia801604.us.archive.org/11/items/bitsavers_xeroxparcteCedarProgrammingEnvironmentAMidtermRepo_13518000/CSL-83-

simply pocket the cash and supply random answers. At the very least, firms want to be sure that readers really do value the book. This will probably require charging readers a modest co-pay.

The controversial alternative is to salt recommendations with speculative suggestions that have nothing to do with the recipient's known preferences.[157] The trouble, of course, is that users cannot be told that the recommender is guessing since its suggestions are bound to be inferior on average. Even so, the occasional correct guesses could still justify the activity by (a) opening new genres to the reader herself, or else (b) identifying new quality titles which can be shared with a wider audience. The former should be uncontroversial since we expect profit-maximizing firms to stop when further guesses would make the recipient worse off in expectation. However, there is a danger in the latter case that companies could use some customers as guinea pigs for others. This might or might not be an acceptable extension of the tithing principle, always assuming that the burdens and benefits are equally shared on average.

Making Word-of-Mouth Predictable. Today's publishing institutions evolved to reduce the blind luck that lets some books find early, sympathetic audiences while other, equally deserving titles do not. The problem, historically, was that there was no practical way to monitor or target word-of-mouth networks in physical space. This limited interventions to expensive and comparatively ineffective mass advertising campaigns. The difference in the digital age is that word-of-mouth is often visible on-line. This potentially lets publishers create a custom network for each new title and then hand-pick the first generation of readers followed by a second and then a third group if the response is positive. Potentially, at least, this would remove the randomness of word-of-mouth markets so that every title is exposed to its ideal audience. Meanwhile, the use of narrowly targeted communications will greatly reduce marketing costs. Web advertising already tailors its ads to users, and Big Data will make this still better over time.

C. *Welfare*

Vertical integration would immediately fix most of the pathologies associated with middlemen[158] while leaving the deeper problem of IP-pricing in place. Subscription models would, however, ameliorate the latter by cutting readers' search costs and enforcing the Hemingway

11_The_Cedar_Programming_Environment_A_Midterm_Report_and_Examination.pdf (explaining how Xerox encouraged its employees to explore technologies by providing a "computationally rich environment" that the general public would not enjoy for years).

157 Netflix already does this. *See* O'Reilly, *supra* note 28.

158 *See supra* "Class 2" items described at Table 1.

principle that old and new books should compete on a level playing field. In the longer term, the rise of electronic paper will similarly eliminate incentives to overproduce physical copies knowing that forty percent will eventually be "pulped."

Potentially, at least, the biggest efficiency gains depend on price discrimination. The basic philosophical question is whether efficiency – in this case increased readership – is worth having when it systematically enriches sellers compared to readers. We return to this in Section IX. In the meantime, we note two subsidiary issues. First, the policy judgment would be much easier if we knew that sellers planned to reinvest their profits in expanded search. Better empirical studies could shed important light on this issue. Second, details matter. In the near-term, publishers will almost certainly depend on price discrimination schemes in which readers who set a high value on texts buy hardbacks instead of e-books.[159] While this kind of "second degree" scheme usually improves welfare, scholars have identified enough counterexamples to make us hesitate. This caveat will disappear if and when publishers transition to "first degree" schemes that set prices according to each customer's individual tastes.

VI. THIRD FUTURE: OPEN SEARCH

Big Data will accelerate vertical integration. But it is still not clear who will benefit. In the meantime, publishers are surely afraid that Amazon will charge much more for search than they will ever receive in revenue from an expanded market. Many readers will have seen this movie before. Silicon Valley companies facing similar threats have joined forces to create open source alternatives for at least twenty years now.[160] Small wonder that the Big Five have similarly discussed launching a joint on-line digital market to challenge Amazon.[161] Once this happens, search would follow naturally. The only question is what institutions are best-suited to house it.

---

[159] *See generally*, Michael J. Meurer, *Copyright Law and Price Discrimination*, 23 CARDOZO L. REV. 55 (2001).

[160] Familiar examples include Apache (enterprise solutions), embedded LINUX operating systems for consumer durables, and the IBM-funded Eclipse collaboration (consulting industry developer tools). *See, e.g.*, Maurer, *supra* note 148; Joachim Henkel, *Software Development in Embedded Linux — Informal Collaboration of Competing Firms*, *in* WIRTSCHAFTSINFORMATIK PROCEEDINGS 2003/BAND III 81, 81–89 (2003).

[161] U.S. v. Apple, Inc., 791 F.3d 290, 300 (2d Cir. 2014) (publishers considered various joint strategies to defeat Amazon, including "possibly creating an alternative ebook platform"). Individual Big Five publishers are too small to do the job alone. Publisher Barnes & Noble's forays into on-line retailing have been perennially troubled. *See* Jeffrey A. Trachtenberg & Michael Calia, *Barnes & Noble to Keep Nook Business*, WALL STREET J. (Feb. 26 2015), https://www.wsj.com/articles/barnes-noble-to-spin-off-college-books-business-1424960675 .

A. *Joint Ventures*

The most conventional solution would be for publishers to provide search services through a commercial joint venture. This would raise familiar antitrust concerns to the extent that the Big Five organized the project in ways that excluded smaller rivals or used search fees to support high book prices.[162] We return to these issues in Section IX.B. In the meantime, the more immediate problem is that the last great wave of joint ventures in the Eighties was uniformly disappointing. The reason seems to be that each company tried to steer R&D in directions that favored its own business. The resulting fights dissipated most if not all of the gains from sharing.[163]

B. *Crowd-Sourcing*

Joint ventures used to be the end of the story. Today, however, the New Economy offers various non-proprietary ("open") alternatives. The simplest of these is "crowd-sourcing" in which a community voluntarily performs some on-line task for its own benefit.[164] Perhaps the biggest advantage is that replacing cash payments with volunteers greatly reduces the conflicts of interest that crippled joint ventures in the past.[165]

The downside of crowd-sourcing is that the number of volunteers is essentially fixed. While they might supply enough effort, coordination, and hierarchy to solve the problem, this can only happen by accident, and without a price signal there is no obvious way to correct shortfalls. That said, improved architectures can sometimes make the existing effort go further. Here, the Web plausibly offers several advantages over earlier face-to-face methods. First, it expands the geographic pool of volunteers from audiences sitting around a campfire to the entire world. Second, we have argued that unpaid volunteers have little appetite for taking orders. But they may sometimes be indifferent across several tasks, allowing computers to prioritize those most valuable to the collaboration. This would be particularly useful for ensuring that volunteers review many different

---

[162] Maurer & Scotchmer, *The Essential Facilities Doctrine: The Lost Message of Terminal Railroad*, 5 CALIF. L. REV. CIR. 287 (2014).

[163] Maurer, *supra* note 148.

[164] Conventional definitions of crowd-sourcing include "1. an organization that has a task it needs performed, 2. a community (crowd) that is willing to perform the task voluntarily, 3. an online environment that allows the work to take place and the community to interact with the organization, and 4. mutual benefit for the organization and the community." DAREN C. BRABHAM, CROWDSOURCING 3 (2013). This does not necessarily exclude institutional architectures that impose coordination and hierarchy, though these are seldom emphasized. We address more structured organizations under the text accompanying *infra*, notes 166 through 170.

[165] RICHARD MORRIS TITMUSS, THE GIFT RELATIONSHIP: FROM HUMAN BLOOD TO SOCIAL POLICY (1972).

titles instead of returning to the most familiar ones over and over again.

Against these benefits, web methods also introduce a new problem. Recommendation engines are mostly useful when they provide the kind of clear, determinate advice that human booksellers provide. Asking readers to wade through thousands of Goodreads-style book reviews falls far short of this standard. This brings us back to the problem of merger. It may be that Big Data can accomplish this final step algorithmically without human intervention. But, if it does not, organizers will need to invent some analog to the ancient bards who reduced disparate audience suggestions to a single definitive product.

C. *Traditional Open Source*

Conventional open source collaborations go beyond crowd-sourcing by adding an inner kernel of developers who enforce the group's quality standards, set priorities across projects, and minimize duplication.[166] Like crowd-sourcing, there is no guarantee that open source can deliver enough effort. Whether it does or not depends on two separate channels:

> *Basal Rate.* Open source enthusiasts are sometimes motivated by intrinsic rewards like altruism, desire for reputation, and the pleasure of creating things. The supply of these is fixed and depends, among other things, on how many similar projects have solicited volunteers already.
>
> *Material Incentives.* Open source enthusiasts frequently act for material reasons like building tools for their own use, learning commercially salable skills, and demonstrating skills to future employers. While these are an imperfect stand-in for the price signal,[167] they can nevertheless nudge supply and demand into closer alignment. The question is how many such incentives exist. Certainly, the number of Americans hoping to land commercial editing jobs must be tiny. At the same time, own-use incentives are potentially more promising, since improved search would at least deliver better titles to volunteers. We return to this in Section VII.C.

The scale of the effort is daunting. A really complete search would require tens of millions of American to read and review one additional

---

166 On leadership in open source, see Maurer & Scotchmer, *supra* note 75 at 305–06. Large volunteer databases in academic science similarly depend on a small central group of editors. Probably the best known example is academic physics. *See About the Particle Data Group*, PARTICLE DATA GROUP, http://pdg.lbl.gov/2017/html/about_pdg.html (last visited Apr. 12, 2018). The hierarchy advantages of open source are ultimately contingent on volunteers' motives to supply labor and, at least implicitly, put up with taking orders and tackling what will often be second-choice projects.

167 Coding skills can be demonstrated in many ways. The large excess of open source developer tool projects compared to, say, accounting packages shows that "what software engineers like" is only distantly related to "what consumers value."

book each year.[168] By comparison, today's biggest open source collaboration only mobilize effort from a few thousand developers, each of whom invests weeks to months of concentrated effort per year.[169] The analogy is somewhat more favorable if we include the larger group of casual volunteers who occasionally report bugs .[170] Even by this relaxed standard, however, an open search collaboration would still require many more volunteers than any existing software project.

D. *Commercial Open Source*

The discussion so far suggests that conventional open source will have trouble eliciting enough effort to feed Big Data's appetite for information. Software companies facing similar shortfalls routinely try to fill the deficit by donating code, equipment, cash, and encouraging employees to "volunteer" on company time. The problem with the last option is that it reintroduces the same conflicts of interest that have traditionally afflicted joint ventures. This, however, is partly offset by the fact that volunteers keep their status in the collaboration after they change employers. That gives them a stake in the collaboration's long-run prosperity, and encourages them to suppress short-sighted strategies that unfairly benefit their current employers over other sponsors.[171]

Probably the biggest danger of commercial support is that paid workers could drive out volunteers so that overall effort might not change. In practice, evaluating this risk forces us to ask why members volunteer in the first place. The most reasonable suggestion, following Bénabou and Tirole,[172] is that volunteers participate to convince themselves that they follow "higher" motives than money, and that

168 We have said that the U.S. produces one million new books each year. Assuming 1,000 reader groups, this naively requires one billion reads, or just over three books per American per year. The actual number would probably be less because the reactions of many reader groups are predictably aligned by genre. Even so, the required number of volunteers would almost certainly run into the tens of millions.

169 The biggest open source projects in 2016 included Firefox (2,367 volunteers contributing 9,132,501 lines of code), Linux Kernel (11,247 volunteers/15,746,046 lines) and Apache (114 volunteers/2,240,171 lines). *See What are the Biggest 3 Open Source Projects by Total Programming Effort?*, QUORA, https://www.quora.com/What-are-the-biggest-3-open-source-projects-by-total-programming-effort (last visited Apr. 12, 2018). Reducing lines of code to man-days of effort is more speculative. The usual commercial estimate is ten lines per developer per day yields. *See* FREDERICK C. BROOKS JR., THE MYTHICAL MAN MONTH: ESSAYS ON SOFTWARE ENGINEERING (1975). This leads to estimates of individual effort that are either large (140 days per year for LINUX) or nonsensical (1,965 days per year for Apache). More realistic estimates would presumably take account of the fact that open source offloads many tasks onto end-users. In any case, the basic picture seems clear: Open software mobilizes far fewer volunteers than a search collaborative would need.

170 Maurer & Scotchmer, *supra* note 162, at 300–01.

171 TITMUSS, *supra* note 165. Economists conventionally assume that open institutions are more transparent than proprietary ones. If so, policymakers should *ceteris paribus* prefer them to the extent that they make antitrust conspiracies less likely.

172 Roland Bénabou & Jean Tirole, *Incentives and Prosocial Behavior*, 96 AMERICAN ECON. REV. 1652, 1654–65 (2004).

commercial support could wreck the illusion. The good news, for now, is that this does not seem to be happening. Most obviously, Amazon's acquisition of Goodreads has not prevented volunteers from posting tens of thousands of reviews.[173] History also helps: The dominant cultural narrative for centuries has invariably lionized paid commercial editors at least as much as amateur literati..

E. *Welfare*

Open search collaborations would let members eliminate wasteful duplication and pool data within a single entity. To the extent that they harvested volunteer labor, they would also eliminate the need for above-cost pricing and open the door to expanded sales. The only question is whether legislators would recognize the possibility and narrow copyright accordingly. If they do not, publishers could easily adapt to shared search by reducing their own efforts while maintaining high prices.

The problem, of course, is that open search might not supply enough effort. Commercial open source models remedy such shortfalls by reintroducing IP incentives. Worse, commercial open source licenses usually guarantee that every company receives exactly the same product. This suppresses incentives to compete and contribute effort in the first place.[174]

Finally, cheaper search would make it easier for readers to find titles. One immediate effect would be to increase purchases from small independent publishers, although the Big Five might still prefer this to an Amazon search monopoly.

VII. CHOOSING TOMORROW'S INSTITUTIONS (A): SURVIVAL-OF-THE-FITTEST

The question remains which of our three futures will emerge. This Section starts from the Darwinian view that the future belongs to whichever institutions supply the most – and also the best mix – of information. Depending on how Big Data's appetites evolve, this could favor proprietary but also open solutions.

A. *Objective Data*

Supplying machine-generated records of consumers' on-line shopping behaviors presents two distinct problems. The first involves information collected *gratis* as a "spinoff" from normal business operations. Here harvesting is mostly a (solved) technology problem so

[173] In the software world, the Eclipse Foundation has similarly managed to combine massive corporate support with significant donated labor. *See* Maurer, *supra* note 148.
[174] For the detailed argument, see Maurer, *supra* note 148.

that choice of institution barely matters: Future institutions will go on harvesting this data whether they are proprietary or open. The more interesting question is what will happen if mature Big Data technologies need more data than the basal rate provides. Here there are two possibilities: Firms can either share data they already possess, or they can invest in developing new data that would not exist otherwise. Proprietary models and open methods have radically different strengths along these dimensions.

Proprietary Models. The argument for additional IP rights holds that letting companies resell data will encourage them to acquire more information in the first place. Doctrinally, however, it is not clear how existing incentives could be expanded. Trade secret law already provides protections for bilateral and multilateral licensing between corporations, and statutory protection for published data would serve no obvious purpose.

In the meantime, stronger IP rights would be costly. Once the data exist, policymakers should want society to use them as much as possible. But in that case the correct policy is to price them at zero. The problem with trying to offer higher prices for new data is there is no obvious bright line standard that can prevent them from attaching to preexisting spinoff data as well.[175]

Open Search. Unlike IP, open institutions maximize sharing by setting the price of data to zero. But this is only beneficial to the extent that corporations find reasons to share in the first place. This decision is highly fact dependent. Suppose that sharing occurs and improves search. The positive side for publishers is that this will grow the overall market, while making their own titles more visible. However, this is offset by the expectation that better search will also help readers find possibly superior titles from competing publishers. This last factor will often be substantial in today's highly imperfect book markets, where sales often depend on private customer lists.[176] The good news is that the balance is subject to a tipping dynamic, so that the profit calculus will change over time to the extent that a successful site becomes a go-to resource for readers.

---

[175] Experience with the European Database Directive is instructive: While the framers originally tried to limit its protection to data that had been collected at "substantial" effort, the absence of a clear standard ensured that practically all commercial data would qualify. *See* Stephen M. Maurer, P. Bernt Hugenholtz & Harlan Onsrud, *Europe's Database Experiment,* 294 SCIENCE 789, 789–90 (2001).

[176] *See* Morris Rosenthal, *Questions About Books Sales: How Many Copies Did My Book Sell?*, FONER BOOKS, http://www.fonerbooks.com/q_sales.htm (last visited Apr. 26, 2018) ("[m]any successful small publishers don't do well on Amazon, they primarily succeed through aggressive marketing to niche audiences through direct marketing.").

B. *Simple Judgments*

Simple judgments present the closest economic analog to "normal" industrial research like screening drug compounds. On the one hand, individual assessments are cheap: Deciding whether a title is "happy" or "sad," for example, only takes a few minutes. On the other, search would require a great many assessments – potentially up to several million per year. Proprietary and open models present very different strategies for mobilizing this effort.

Proprietary Models. The great advantage of traditional IP is tithing, *i.e.* spreading the cost of one-time discoveries across thousands of consumers. Theory alone cannot say whether this familiar model can be stretched to supply Big Data's thirst for information. Even so, the precedents are encouraging. After all, characterizing a million books is not very different from Thomas Edison's evaluation of thousands of candidate light bulb filaments in the 19th Century[177] or Big Pharma's testing of tens of thousands of compounds for each successful drug in the 20th Century.[178] Granted that book search is more daunting – the number of candidates is perhaps 100 times larger while profit margins are thinner – these factors are at least partly offset by the fact that workers would only need a minute or two to process each book.[179]

Traditional Open Source. Because open source relies on unpaid volunteers, it cannot make the cross-payments that tithing requires. This limits effort to whatever direct benefits volunteers derive from participating. Probably the closest analog to an open search site is Project Gutenberg, whose volunteer members have identified, digitized, and proofed 54,000 public domain titles over the past two decades.[180] Unfortunately, this number is still 10,000 times smaller than what would be needed to assign attributes to one million books each year, although this is at least arguably offset by the fact that evaluating new titles would be more attractive and less burdensome than digitizing pre-1923 titles.

In the meantime, theory is encouraging. First, open collaborations work best where the work can be divided into small ("granular") packets.[181] This is maximally satisfied for simple judgments. Second,

---

177 *See* Consolidated Electric Light Co. v. McKeesport Light Co., 159 U.S. 465 (1895).

178 Solomon Nwaka & Robert G. Ridley, *Virtual Drug Discovery and Development for Neglected Diseases Through Public–Private Partnerships*, 2 NATURE REVS.: DRUG DISCOVERY 919, 919–20 (2003).

179 We assume that assigning attributes is the quintessential example of "judging a book by its cover" requiring, at most, a minute or so for workers to skim chapter headings and text.

180 *Free e-books - Project Gutenberg*, GUTENBERG, https://www.gutenberg.org/ (last visited Apr. 12, 2018).

181 Benkler, *supra* note 51. The idea of a single threshold is admittedly simplistic. Volunteers in classical open source collaborations have enormously different appetites for work. Tiny (< 10%) minorities often contribute the bulk (> 75%) of all code. *See, e.g.*, Maurer & Scotchmer, *supra* note 162, at 300–01.

volunteers who donate judgments can expect better predictions (and reading enjoyment) in return. This should lead to greater effort, although the effect would be heavily suppressed by the usual free-rider dynamics in which each volunteer waits for someone else to do the job first. While we should probably expect somewhat less effort than proprietary institutions, open source solutions remain plausible.[182]

Commercial Open Source. Corporations often find it in their business interest to bolster traditional, volunteer-based collaborations with cash donations, in-kind contributions, and paid manpower. Since these resources must be paid for, this usually leads to a backdoor reintroduction of IP, tithing and above-cost prices. Even so, commercial open source could still be less distortionary than traditional proprietary models to the extent that it mobilizes volunteer labor and avoids the pathologies that handicap traditional joint ventures.

C. *Complex Judgments*

This category includes human editor judgment and its lay reader analogs. The main feature of this data is that reactions can and usually do vary from one reader to the next. This implies that these data almost always shed light on *both* the title *and* the reader simultaneously. While it is analytically convenient to discuss each effect separately, readers should keep in mind that real world data almost always present both kinds of information within a single indivisible package.

Problem 1: Evaluating Titles. We have argued that judging quality would require Americans to read and evaluate roughly one million books per year. This goal is not very different from our simple judgments discussion and suggests that traditional IP strategies based on tithing have a reasonable chance of success.

The question is whether open source models can match this effort. This would require expanding Goodreads' current effort by 3,000 times. Intuitively, at least, this benchmark seems sufficiently daunting to give proprietary institutions the edge. The catch, as we have said, is that "Evaluating Titles" is only part of the problem. Before reaching any definite conclusions, we must also consider the second and still-harder task of "Diagnosing Readers."

Problem 2: Diagnosing Readers. We have argued that mature Big

[182] The standard analysis holds that players adopt "game of chicken" strategies in which each delays making a contribution with hopes that somebody else will do the job first. The silver lining is that games of chicken cannot go on indefinitely so long as one or more players really wants the good. This leads to "mixed strategies" in which members randomly decide whether or not to work in each period. *See generally*, Maurer and Scotchmer, *supra* note 162 (literature review). Participation is further boosted where players have heterogeneous needs so that each self-selects into projects that she values more than anyone else. *See*, Joachim Henkel, *The Jukebox Model of Innovation – A Model of Commercial Open Source Development*, (Ctr. for Econ. Res., Discussion Paper No. 4507, 2004).

Data techniques would require readers to provide opinions for roughly fifty books. This would be easy for top tier readers and hopeless for those near the bottom. For everyone else, the accuracy of predictions would depend on how well proprietary and open institutions expand the dataset by persuading customers to read more and/or different books than they normally would.

This, however, exposes a problem that IP normally sweeps under the rug. The simplest and most general economic models predict that copyright divides the social value of reading equally between buyers and sellers.[183] It follows that an investment in search that grows the market by $1 will return just 50¢ to each side. But in that case, *neither side should ever invest for less than a two hundred percent return.* In principle, publishers and readers could evade this limit by making reciprocal promises to invest effort. This, however, would require millions of (unenforceable) bilateral contracts between publishers and readers. This is plainly impractical.

The only other possibility is to change the economics so that one side makes all the investments and receives all the benefits. We have already seen how publishers could do this through price discrimination. This, however, would imply a massive revenue transfer to publishers, although we might not care if we were confident that the profits would be reinvested in search. The analysis for open institutions is symmetrical. Since volunteer-driven search costs nothing, IP markups would no longer be necessary. Getting rid of them would then increase the value of reading and, implicitly, readers' incentives to volunteer in the first place. In the meantime, the economics of diagnosing readers flips the usual arguments for proprietary solutions compared to open source. On the one hand, tithing no longer applies: Once the goal becomes profiling each and every reader, the idea of spreading R&D costs across some larger population no longer makes sense. On the other hand, there is also no incentive to free-ride: Readers who fail to report their opinions know that no one else can do the job for them.[184] This immediately eliminates the deepest obstacle to traditional open source. We are left with the surprising intuition that open institutions will often

[183] By "simple" we mean the result is restricted to models that assume (a) a linear demand curve, and (b) that the monopolist charges the same price to every customer. *See, e.g.*, Stephen M. Maurer & Suzanne Scotchmer, *Procuring Knowledge*, 15 ADVANCES IN THE STUDY OF ENTREPRENEURSHIP, INNOVATION, AND ECON. GROWTH 1, 31 (2004).

[184] This motive is particularly salient when on-line sites ask customers to correct incorrect assumptions. *See* Daniel Tunkelang, *The Napoleon Dynamite Problem*, THE NOISY CHANNEL (Nov. 21, 2008), http://thenoisychannel.com/2008/11/21/the-napoleon-dynamite-problem (noting how sites ask users to rate their recommendations, or else correct estimation errors when, for example, vendors mistake a one-time gift purchase as evidence of the customer's own taste); *see also*, Joseph A. Konstan & John Riedl, *Deconstructing Recommender Systems: How Amazon and Netflix Predict Your Preferences and Prod You to Purchase*, IEEE SPECTRUM (Sept. 24, 2012), http://spectrum.ieee.org/computing/software/deconstructing-recommender-systems.

diagnose readers more effectively than proprietary ones.

But that is theory. We have assumed a tacit bargain in which readers who take over search effort receive lower markups in return. But the IP monopoly is set by statute and will continue in its present form until Congress decides to change it. Ideally, Congress would notice that readers are donating more to search and narrow the statute accordingly. If not, we argue below that publishers can achieve a similar (if second-best) result informally.

VIII. CHOOSING TOMORROW'S INSTITUTIONS (B): PATH DEPENDENCE

So far, we have made the classical Darwinian assumption that the most efficient institutions will win out. But modern biologists and economists know that established institutions often outlast their shelf lives.[185] This Section explores various mechanisms that could lock society into inferior choices.

A. *Patents*

So far we have treated Big Data as inevitable, ignoring how the practicalities of funding could favor some players over others. This is reasonable for publishing, where basic search technologies will usually be inherited from more lucrative markets like movies and games. When this happens, we expect IP owners to license whichever institutions deliver the most value to readers, since these also tend to generate the most royalties. We now explore more complicated scenarios where this logic could go astray.

Leveraging. We have assumed that the IP owner's only asset is the statutory monopoly. However, owners might possess enough preexisting market power to build a broader or longer-lasting monopoly than the statute provides. This concern would be especially strong if a dominant firm like Amazon were to invest in R&D or else buy up existing patents. At the same time, policymakers should recognize that publishing disfavors such strategies. We have argued that merchants' ability to charge high margins for digital books is already capped by the threat of piracy. If leveraging is possible at all, the resulting markups will likely be limited.

Patent Thickets. Judging from history, there is a strong chance that Big Data will advance through clouds of small improvements.[186] This

185 See, e.g., STEVEN J. GOULD, PUNCTUATED EQUILIBRIUM (2007) (persistence of less adapted species in biology); Paul A. David, *Why are Institutions the 'Carriers of History'? Path Dependence and the Evolution of Conventions, Organizations and Institutions*, 5 STRUCTURAL CHANGE & ECON. DYNAMICS 205 (1994) (discussing the role of historical accident in shaping institutions).

186 The fact that successful algorithms are often poorly understood even by the inventors

could raise several problems:

> *Overpatenting.* We normally associate patent rewards with royalties. However, large firms in the electronics industry sometimes care more about deterrence strategies in which each player amasses patents to threaten its rivals. Since these threats tend to cancel, the resulting arms race can continue indefinitely. This wastes resources by forcing R&D deep into diminishing returns.
>
> *Transaction Costs.* Individual patents can offer so little value that royalty agreements are not worth negotiating. At this point, IP incentives become ineffective. Finding ways to reduce transactions costs restores incentives for socially useful R&D.
>
> *Anticommons Issues.*[187] Negotiations sometimes fail even when agreement would generate profit for all concerned. While the reasons for deadlock are surprisingly obscure,[188] the effect probably scales with the number of patents.

The familiar cure for all three problems is to allow patent pools that batch-license IP. The trick is to do this without restricting entry and cartelizing prices. We return this issue in Section IX.

Overpayment. At its most basic level, patent law trades faster innovation for high prices and reduced consumption. We should, therefore, worry that cheap and easy inventions will be over-rewarded. This will be hard for judges to fix, since reducing reward after the fact makes future incentives less credible.[189] Government policymakers, on the other hand, can do the next best thing by making sure that obvious R&D projects are funded outside the patent system. The approach is particularly promising for Big Data technologies, where good ideas have almost always been scarcer than the comparatively small grants and prizes needed to test them.[190]

---

themselves further argues against the possibility that some "fundamental" or "pioneer" patent will eventually emerge to dominate the field.

187 Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCIENCE 698, 698−99 (1998).

188 *Id.* Some explanations invoke conventional economic models, most notably when parties refuse licenses in hopes of getting a better offer. Others argue from incentives that have nothing to do with profit maximization, including a desire to cripple rivals. Finally, scholars argue that negotiators can sometimes suffer from "cognitive biases" that lead them to systematically overvalue their patents. *Id.*

189 But see the disputed doctrine that courts should accord "pioneer patents" greater breadth than lesser, but still patentable discoveries. John R. Thomas, *The Question Concerning Patent Law and Pioneer Inventions*, 10 BERKELEY HIGH TECH. L. J. 35 (1995); Michael J. Meurer & Craig A. Nard, *Invention, Refinement and Patent Claim Scope: A New Perspective on the Doctrine of Equivalents*, 93 GEORGETOWN L. J. 1947, 1989 (2005).

190 The scheme assumes that agencies will require recipients to renounce their IP rights. The economic case for such policies can be found in, for example, SUZANNE SCOTCHMER, INNOVATION AND INCENTIVES 242, 242−47 (2004).

B. *Copyright*

Modern copyright anomalously covers *both* literary works *and* software programs.[191] Big Data creates issues for both categories.

Copyright in Software. Big Data methods only become useful when they are reduced to working code. Copyright famously protects each of these implementations as a separate "expression" of the same underlying "idea."[192] This invites would-be competitors to continue writing duplicative software until there is no longer enough profit to fund yet another search program.

The question is what copyright can do about the problem. Even if copyright disappeared entirely, firms could usually achieve similar protections by keeping software in-house as a trade secret. No imaginable reform is likely to change this.[193] The prospects for reforming copyright in open collaborations are more promising. While so-called "copyleft" licenses may sometimes be needed to suppress free-riding, most collaborations seem to get along fine without it.[194] Antitrust judges should view such licenses skeptically.

Copyright in Content. We have argued that IP policy should balance the benefits of search against high book prices. But Big Data has changed both sides of this equation, and will probably change them even more in the future. In the short run, this suggests that the recent erosion of digital book prices may have gone too far. In principle, at least, stronger copyright could encourage publishers to invest more in search so that book sales expand faster than high prices suppress them. In the long run, however, we have emphasized that a mature Big Data technology will need massive amounts of self-reported data. High book prices could limit consumers' incentives to supply this information, implying that Congress would be wiser to dilute IP instead. We return to this topic Section IX.

C. *Trade Secret and Database Rights*

We have emphasized that Big Data capabilities ultimately depend on access to a rich and varied supply of information. This suggests that dominant players can cripple rivals by withholding data, potentially inviting smaller firms and open collaborations to retaliate in kind. At this point, the publishing industry could become so balkanized that Big Data might never reach its full potential.

Whether dominant firms really do decide to hoard data will turn on two calculations:

---

[191] *See* 17 U.S.C. § 102(a); Apple Computer, Inc. v. Franklin Computer Corp., 714 F.2d 1240 (3d Cir. 1983).

[192] *See* Computer Associates International, Inc. v. Altai, Inc., 982 F.2d 693 (2d Cir. 1992).

[193] *See infra* Section IX.B (discussing compulsory licensing).

[194] Maurer, *supra* note 148.

> *Deterrence.* Forcing challengers to replicate hoarded information could make entry unaffordable, reducing competition and allowing incumbents to raise prices.[195] This logic is especially strong in the publishing industry, where the costs of providing data will often be a large fraction of the overall social benefits.
>
> *Strategic Sharing.* If new entrants cannot be deterred, dominant firms must still decide whether hoarding will increase their profits. The answer will normally be "yes" where (a) hoarding translates into a significant advantage in search and therefore sales, and (b) this advantage is larger than the expanded sales that every firm would receive by letting pooled data grow the overall market.

Finally, policymakers should also worry about the symmetric case where new entrants decide to hoard data.[196] The best case for allowing such a right would be if policymakers thought that the newcomers would use withholding as a bargaining chip to force bilateral sharing. Open collaborations may also need viral restrictions to stabilize their memberships against free-riding, though here the empirical evidence is obscure and judges would be wise to be skeptical.[197]

The problem in both cases is imagining how lawmakers could modify trade secret law short of mandated sharing. We return to this possibility in Section IX.B.

IX. MANAGING THE REVOLUTION: LAW AND POLICY

We have argued that Big Data opens the door to more efficient institutions. But we have also identified roadblocks that could block the transition indefinitely. Judges and policy makers should design rules that let better institutions emerge if and when Big Data makes them possible.

A. *Reforming Apple*

Scholars usually imagine the law in constant flux. But in a small market like publishing, landmark cases are so few and far between that evolution proceeds through what biologists call "punctuated equilibrium," *i.e.* long periods of stasis that are only occasionally interrupted by change.[198] In the normal course, we can expect the Second

---

195 The barrier is limited by the fact that reader tastes change and new books are constantly being published. This suggests that the "shelf life" for data cannot be much more than a decade or so.

196 GPL open source licenses already prohibit sharing in the software context. I have argued elsewhere that this is, or at least ought to be, illegal under the Sherman Act. *See* Maurer, *supra* note 162.

197 *Id.*

198 *Punctuated equilibrium*, PBS, http://www.pbs.org/wgbh/evolution/library/03/5/l_035_01.html (last visited Apr. 25, 2018).

Circuit's *Apple* decision to frame industry practice for decades.[199] The only question is whether judges should find some special reason to revisit it sooner. We have argued that the rise of e-books offers a once-in-a-generation chance to restore the kind of strong price discrimination that cheap paperbacks provided fifty years ago. Since the old system sometimes expanded readership by nearly an order of magnitude, judges should think carefully before refusing the gift.

Apple should have been the perfect case to decide whether price discrimination was in society's interest and, if so, when and how authorities should manage it. Instead, the parties persuaded the Second Circuit to restrict its analysis to just one market – e-books – while ignoring the parallel impacts to e-reader tablets and hardbacks.[200] This turned price discrimination into a non-issue.

Background. The court's findings are quickly stated. Four large publishers conspired with Apple to take e-book pricing away from Amazon.[201] Crucially, each member of the triangle sought to use this power differently. Amazon saw cheap e-books as a loss-leader for building a standard that would allow it to dominate e-readers.[202] Apple had the opposite strategy: It wanted to keep e-reader prices high, and therefore allied itself with publishers to stop Amazon from starting a price war in cheap content.[203] Finally, the big publishers were indifferent to e-reader sales, but wanted to keep low e-book prices from eroding their hardback profits.[204]

Had the conspiracy succeeded, publishers would have used their power to raise prices. The Second Circuit reasoned that this was identical to price fixing.[205] But this ignored that the conspiracy would have ended well before any prices were set. Furthermore, each publisher would have made this latter judgment independently.[206] Doctrinally, at least, this raises the question whether this last-minute competition was enough to break the causality between the publishers' conspiracy and

[199] U.S. v. Apple, Inc., 791 F.3d 290, 311 (2d Cir. 2014).

[200] Formally, the court adopted the parties' stipulation that the case would be limited to the pricing of "trade e-books" while excluding "e-readers" and hardbacks. *See Apple, Inc.*, 791 F.3d at 311.

[201] The mechanics of the scheme were famously involved and included the use of a "Most Favored Nations" clause that would have forced the Big Five to match Amazon's e-book prices for Apple users. *Id.* at 305. This had two effects. First, publishers that let Amazon subsidize book prices would have had to give Apple matching discounts from their own pockets. Second, the clause ensured that Amazon could never offer cheaper e-books than Apple no matter how much it spent. This eliminated its business case for offering below-cost titles in the first place.

[202] *Id.* at 299.

[203] *Id.* at 340.

[204] *Id.* at 300, 305.

[205] *Id.* at 328.

[206] Realizing that the Big Five would surely raise prices, Apple also demanded that publishers cap prices *below* the new market equilibrium. *Id.* at 317.

the claimed antitrust injury (higher prices) that followed.[207]

Normally, one might dismiss such theories as wordplay. But the distinction between pre-market agreements and subsequent competition is extremely substantive. Sherman Act judges have long recognized that many if not most industries are impossible without common ground rules, and that the parties should be allowed to set them when the alternative is no market at all.[208] Furthermore, it is hard to imagine a rule more basic than who should set prices.[209] But in that case, the only remaining issue is whether the publishers' purpose to price discriminate takes their agreement outside the safe harbor for agreed market structures. This raises the surprisingly obscure question of whether price discrimination is consistent with the Sherman Act.

Is Price Discrimination Desirable? This article has invoked the familiar freshman economics argument that price discrimination is *ipso facto* desirable because it reduces monopoly distortion and increases total welfare. Remarkably, there is almost no case law on whether the Sherman Act endorses this result.[210] To the contrary: Many legal scholars argue on historical grounds that Congress intended antitrust as a consumer protection statute.[211] If so, welfare impacts are irrelevant, and price discrimination should be *ipso facto* condemned because it enriches producers over buyers.[212] The counterargument, presumably, is that the Supreme Court later introduced conventional economic

[207] Tort law routinely recognizes that proximate cause is interrupted by independent intervening acts. In this language, the publishers' independent pricing decisions interrupted the causation connecting the antitrust conspiracy to the eventual antitrust injury of higher prices. *See, e.g.*, Derdiarian v. Felix Contracting Corp., 51 N.Y.2d 308, 434 N.Y.S.2d 166 (NY Ct. App, 1980); McCoy v. American Suzuki Motor Corp., 136 Wash.2d 350, 961 P.2d 952 (Sup Ct. Wash 1998).

[208] NCAA v. Board of Regents, 468 U.S. 85, 101−102 (Sherman Act permits self-regulation where " . . . the integrity of the product cannot be preserved except by mutual agreement.").

[209] *Cf.* Leegin Creative Leather Products, Inc. v. PSKS, Inc., 551 U.S. 877 (2007) (Sherman Act permits manufacturers to control of retail pricing subject to rule of reason).

[210] The reason seems to be that courts often encounter price discrimination in predatory pricing contexts that have nothing to do with the classical tradeoff between expanding output and impoverishing consumers. Daniel J. Gifford & Robert T. Kudrle, *The Law and Economics of Price Discrimination in Modern Economies: Time for Reconciliation?*, 43 U. CAL. DAVIS L. REV. 1235, 1293 (2010) ("Dominant" view in both the US and Europe is that total welfare rather than consumer welfare should drive antitrust analysis); Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2, 14−15 (1984) (arguing that price discrimination "can increase the social costs of market power by facilitating price discrimination, thereby increasing monopoly profits over what they would be absent the tie"); Herbert Hovenkamp, *Implementing Antitrust's Welfare Goals*, 81 FORDHAM L. REV. 2471, 2474 ("[F]ew if any decisions have turned on the difference" between consumer surplus and social surplus).

[211] Probably the best argument for this interpretation is that economic efficiency arguments barely existed when Congress passed the Sherman Act in 1890. Indeed, conventional accounts usually hold that modern microeconomic theory dates from the publication of Alfred Marshall's widely-influential book that same year. *See* ALFRED MARSHALL, PRINCIPLES OF ECONOMICS (1890).

[212] Roger D. Blair & D. Daniel Sokol, *Welfare Standards in US and EU Antitrust Enforcement*, 81 FORDHAM LAW REVIEW 2497, 2499 (2013) ("Dominant" view in both the US and Europe is that total welfare rather than consumer welfare should drive antitrust analysis).

arguments to rationalize the statute in *Standard Oil Co. of New Jersey v. U.S.*,[213] and that economic interpretations have grown steadily more important since then. Suffice to say that there is no consensus on the point,[214] and that scholars who prefer efficiency interpretations are nearly as numerous as those who oppose them.[215]

The confusion only deepens when we recall that antitrust law is just part of the puzzle, and that, when we turn to IP law, the policy focus shifts to making sure that inventors receive enough reward to incentivize R&D. Doctrinally, however, IP law is nothing more than the power to exclude. How well or badly this power is actually monetized depends on which business models the Sherman Act allows.[216]

The first step in reconciling IP with competitions policy is to find some common set of goals, at least where the two frameworks overlap. Unlike the Sherman Act, however, no one has ever accused IP law of being a consumer protection statute.[217] To the contrary: judges in patent cases routinely invoke standard microeconomic arguments that praise price discrimination as a useful lever to soften the innovation/IP monopoly tradeoff.[218] The simplest way to do this doctrinally is to find that price discrimination is inherent in the patent grant,[219] so that any Sherman Act objections no longer matter. The argument is even

[213] Standard Oil Co. of New Jersey v. U.S., 221 U.S. 1 (1911).

[214] Jonathan B. Baker, *Economics and Politics: Perspectives on the Goals and Future of Antitrust*, 81 FORDHAM L. REV. 2175, 2177 (2013). Despite decades of trying, the antitrust community has still not reached a "durable consensus" over the economic goal that antitrust enforcement should pursue. Indeed, Baker claims that this is probably "asking too much." *Id.* at 2180.

[215] *See, e.g.*, John B. Kirkwood, *The Essence of Antitrust: Protecting Consumers and Small Suppliers from Anticompetitive Conduct*, 81 FORDHAM LAW REVIEW 2425, 2428–30 (2013) (arguing that consumer surplus standard "commands wider support than any other" and "is now espoused by a majority of courts."). Canada has explicitly embraced a total welfare approach, albeit in the limited context of mergers. *See* Marc Duhamel, *On the Social Welfare Objectives of Canada's Antitrust Statute*, 3 CANADIAN PUBLIC POLICY XXIX 301, 302 (2003).

[216] More precisely, antitrust needs to share IP's goals in those cases where the two overlap. This would still leave judges free to adopt other, different competition goals where IP is absent. This is at least consistent with occasional assertions that the Sherman Act is a consumer protection statute "for mergers" and other narrowly defined contexts. *See, e.g.*, Duhamel, *supra* note 215, at 302 (limiting Canada's adoption of total welfare approach to merger context).

[217] Meurer, *supra* note 159, at 91−92, n. 7 (arguing that antitrust law is "less important than copyright and contract law in terms of its influence on price discrimination").

[218] ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1450 (7th Cir. 1996) (upholding shrink-wrap license based on need to suppress "arbitrage [that] would break down the price discrimination and drive up the minimum price at which ProCD would sell to anyone."); *In re* Brand Name Prescription Drugs Antitrust Litig., 288 F.3d 1028, 1030−31 (7th Cir. 2002) (Judge Posner noting ubiquity of price discrimination in hardback books, first-run movies, and generic drug manufacturing).

[219] *See* Broadcast Music, Inc. v. Columbia Broad. Sys., Inc., 441 US 1 (1979) (affirming blanket licensing scheme where royalty varied with buyer's revenues); USM Corp. v. SPS Technologies, Inc., 694 F.2d 505, 512–13 (7th Cir. 1982) (Posner, J.) (noting that "there is no antitrust prohibition against a patent owner's using price discrimination to maximize his income from the patent"), *cert. denied*, 103 S. Ct. 2455 (1983); *In re* Independent Serv. Org. Antitrust Litig., 989 F. Supp. 1131, 1139, (D. Kan. 1997) ("A patent holder's right to price its patented products at different prices to different customers is inherent in the patent grant.").

stronger for copyright, where book production is valued not only for its economic benefits but because it promotes culture and politics. This immediately implies that it is ". . .the total quantity sold [and] not the consumer surplus that matters."[220]

Of course, the Second Circuit might still reject these arguments. But that is not the point. Whether or not the legal case for price discrimination prevails, it is both too strong and too consequential to be decided by inattention. We now ask how incorporating price discrimination into *Apple* might change the Court's analysis.

Fixing Apple. Assume, then, that price discrimination is desirable. How should the Sherman Act implement it when several equally competitive market structures are possible? Logically, there are just three choices. First, courts could dictate whichever structure seems best to them. This, however, violates the Sherman Act's founding conceit that markets are wiser than central planners, along with judges' understandable fear that they will be inundated with requests for guidance. Even so, this could still be desirable if we expected markets to consistently reach the wrong result. Second, courts could invoke the familiar doctrine that conspiracies are more dangerous than unilateral behavior, *i.e.* that Section 2 is narrower than Section 1.[221] In that case judges might adopt a rule that lets publishers negotiate with Amazon as individuals but not collectively. But that would lock in the (im-)balance of power between the parties to the point of installing an "Amazon always wins" rule. Finally, judges could let the firms negotiate whatever arrangements they like. This mirrors the usual Sherman Act deference for letting the parties set ground rules, but should only be acceptable if we think that the parties' private goals really will mirror society's interests.

The question, then, is how well private and public incentives to price discriminate are aligned. This turns out to be a two-part inquiry. The first and simplest question is whether private profit-maximization coincides with the public goal of mitigating lost sales ("deadweight loss"). Here it seems obvious that expanded sales will simultaneously increase profit and increase efficiency by enticing customers who would never agree to the full monopoly price. This suggests that we really can trust the parties to set socially efficient e-book prices.[222]

---

220 Meurer, *supra* note 159, at 92.

221 *See* Copperweld Corp. v. Independence Tube Corp., 467 U.S. 752, 768 (1984) (Sherman Act "treat[s] concerted behavior more strictly than unilateral behavior.").

222 Our argument assumes that the Big Five and Amazon were both motivated by price discrimination. However, one could argue that this interpretation is inconsistent with *Apple*'s holding that Amazon wanted to control e-book prices as a "loss-leader" to build a market position that would dominate e-readers indefinitely. *See generally* U.S. v. Apple, Inc., 791 F.3d 290 (2d Cir. 2014). Fortunately, the discrepancy disappears if we think that consumers weigh cheap titles against the estimated costs of an Amazon monopoly, so that the logic of price discrimination

The second issue is whether the law gives parties enough tools to negotiate practical transactions that reach this result. Recall that the basic dilemma in *Apple* is that the parties possessed just one lever (e-book prices) to price discriminate in two separate markets (e-readers, hardbacks).[223] Given the bare knuckles fight between publishers and Amazon, the Second Circuit understandably theorized this situation as a zero-sum fight for "control."[224] Looking back, this presented a false choice. Ultimately, whether the new profits came from e-readers made no difference to Amazon, to publishers, or to the wider society. Instead, the *only* thing that mattered was *total* profits *regardless* of origin. But in that case, the most lucrative solutions might often involve a little of each, *i.e.* compromise pricing designed to establish less-than-perfect discrimination in both markets simultaneously.

The question, of course, is why Amazon and the publishers never tried to do this. The answer, almost certainly, involves appearances. A successful agreement would have required Amazon and the publishers not just to maximize joint profits, but also to divide the proceeds afterwards. But that would have required cross-payments, for example by having the Big Five pay part of their hardback earnings to Amazon. This was bound to draw accusations that Amazon was extending ("leveraging") its power over digital books to other markets.[225] But judges should know better. If price discrimination really is desirable, the Second Circuit should have the courage not only to say so, but to authorize whatever cross-payments are required to implement it.

B. *Antitrust Restrictions on Sharing*

We have argued that Big Data predictions are only as good as the information they feed on. At the same time, it is easy to imagine competitions authorities challenging the required joint ventures and open search collaborations as anticompetitive. Conversely, authorities could decide to force sharing through compulsory licensing and the essential facilities doctrine.

Joint Ventures and Open Search Collaborations. Joint ventures and open collaborations let parties share and reuse data. However, they also offer tempting levers that can be used to suppress competition.

returns. In any case, the added wrinkle is no worse than other situations where antitrust judges are asked to balance short-term competition "in the market" against long-run ("Schumpeterian") competition "for the market."

[223] Our analysis specifically assumes the "second degree" price discrimination methods described in *Apple. See, e.g. Apple, Inc.*, 791 F.3d 290. While we have argued that "first degree" methods might exist in the future, we ignore the possibility in what follows.

[224] *See, e.g. Apple, Inc.*, 791 F.3d at 305.

[225] It is worth noting that large media companies have considered similar arrangement in the past. *See* SMITH & TELANG, *supra* note 30, at 306 (NBC asked Apple to share part of its iPod sales revenues in licensing negotiations in 2007, although Apple refused.).

Given that each member receives the same predictions as every other member, firms cannot benefit by investing more than competitors. This suppresses R&D effort much as an illicit cartel would. The question is how the losses from reduced investment compare to the benefits of sharing data that no single firm could collect on its own. Since the 1980s, Congress and the courts have consistently struck this balance in favor of sharing.[226]

Whether this same rule should be extended to shared search institutions is plainly fact-dependent. However, we have argued that Big Data depends on the amount(s) and type(s) of available data far more than clever algorithms. This suggests that policymakers should normally prioritize sharing over competition.

Compulsory Licensing. We have emphasized that firms may rationally decide to hoard data. The question remains whether antitrust judges should try to overcome this. The practical difficulty is that compulsory licensing requires judges to specify a royalty. This, in turn, forces them to confront such difficult questions as how much the IP owner invested in research, whether a smaller incentive would have sufficed, and whether some other actor could have done the job more cheaply.[227] This has largely deterred courts outside the special case where IP assets cost little or nothing to develop.[228] Fortunately, this is a good fit for publishing, which depends heavily on spinoffs and donations.

Essential Facilities. We have seen that joint ventures are a natural way to pool information. The extreme version of this argument comes when the economies of sharing are so dominant that individual firms cannot compete otherwise. In this case, the Sherman Act – which normally blocks joint production by competitors – can permit and even compel sharing. This normally takes the form of asking whether some shared facility is "essential" to compete in the industry.

The good news in this analysis is that it does not matter whether the Sherman Act is interpreted in terms of economic efficiency or consumer protection. After all, the goal in both cases is to offer consumers the lowest possible prices. Instead, the only question worth asking is whether this can best be done by promoting competition (low markups) or scale economies (low underlying costs). Here the obvious solution – easier said than done – is to adopt whatever mix promises the lowest prices. The late Prof. Scotchmer and I have argued that this is by far the most natural way to rationalize cases descending from the US

226 *See* Maurer, *supra* note 148.

227 *See* SCOTCHMER, *supra* note 190, at 53−58.

228 Most examples of forced sharing to date involve arbitrary strings of digits like telephone numbers or the "Applications Program Interfaces" that programmers need to interact with particular operating systems. Maurer & Scotchmer, *supra* note 183, at 154 (surveying case law).

Supreme Court's *Terminal Railway* decision.[229]

Of course, deciding to allow sharing is only the beginning. Regulators must still set rules that minimize the risk that firms will use the shared facility to cartelize prices or exclude competitors. The danger in the first case is that the facility's user fees will enter firms' marginal costs and, through them, enforce a monopoly price. The genius of *Terminal Railway* was to realize that such tactics only work when the shared facility can later distribute the resulting profits back to members.[230] This suggests that essential facilities should normally be organized as non-profits that cannot pay dividends. The second risk is that incumbents can manipulate membership to exclude competitors. *Terminal Railway* and most later cases address this by requiring open membership unless and until the facility becomes congested. The rule is even simpler for information goods where we expect benefits to grow forever.[231] Finally, it is worth noting that the two goals are in tension: Facilities that charge fixed up-front fees tend to exclude small players, while per-use charges increase the risk of cartelization. The trick for regulators is to find fee structures that balance these risks.

C. *Innovation Policy*

We have argued that Big Data has changed the costs and benefits of search, and that we should expect even larger changes going forward. In a perfect world, this would encourage Congress to revisit how much IP protection is necessary. While the traditional levers mostly revolve around duration and breadth, we can also imagine more novel solutions. These notably include expanding the pool of owners who are authorized to share IP.

Copyright. We have argued that rebalancing IP's costs and benefits could imply stronger rights in the near-term followed by relaxation later. The question is how much freedom Congress has to do this. Given today's digital technologies, stronger copyright laws may not be feasible. Instead, higher prices may simply persuade consumers to spend more time looking for pirated copies.[232] Attempts to weaken copyright, on the other hand, face doctrinal difficulties. On the one hand, the idea/expression framework is already so narrow that it is hard to imagine further changes short of allowing simple, non-literal copying. At this point, copyright would be nearly irrelevant. On the

[229] U.S. v. Terminal R.R. Ass'n, 224 U.S. 383 (1912); *see generally*, Maurer & Scotchmer, *supra* note 162.

[230] *See Terminal R.R.*, 224 U.S. 383.

[231] HAL R. VARIAN & CARL SHAPIRO, INFORMATION RULES: A STRATEGIC GUIDE TO THE NETWORK ECONOMY 179 (1999).

[232] While Congress could potentially extend duration, this is already so long that further increases would have limited impact. More fundamentally, incentives would remain derisory for the vast majority of titles that have no hope of becoming "classics" in any case.

other hand, copyright's formal duration is already enormously longer than the commercial life of most titles.[233] This suggests that Congress would have to cut duration by at least a factor of ten to significantly change incentives. This seems unrealistic.

That said, the stakes are high. We have argued that Big Data will likely evolve in directions that require massive self-reporting from readers. But in that case readers must be rewarded for their effort, and it is easy to see how high copyright-supported book prices could get in the way. Given the formal obstacles to rewriting copyright law, it might be better to strike the trade privately. One obvious way to do that would be to reward readers who self-report with a discount on future digital book purchases.[234] The trouble is that this would invite free-rider problems unless every collaboration member agreed to honor the discount, in apparent conflict with Sherman Act's ban on agreements to "to raise *or lower* prices."[235] Judges will have to decide whether this result is sensible. We have argued that antitrust should not penalize firms for agreeing to one competitive market structure over another. That reasoning is even stronger here, where lower prices would simultaneously improve search and increase demand.

Data Portability Statutes. We have argued that commercial firms often have good reason to hoard data. However, a recent EU directive has largely taken this option off the table by requiring member states to pass "data portability" laws that let consumers to force data sharing among firms:

> The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller . . . and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided.[236]

At first glance, data portability looks like a still more ambitious version of compulsory licenses, this time imposed as a blanket judgment

---

[233] The current default duration expires seventy years after the author's death. By comparison, commercial publishers typically offer titles for a decade or so.

[234] The scheme would not be perfect, since profit-maximizing publishers will normally limit discounts so that they preserve some monopoly profit on each book sold. But this partial fix is still better than no improvement at all.

[235] U.S. v. Socony-Vacuum Oil Co., 310 U.S. 150 (1940) (emphasis supplied).

[236] Commission Regulation 2016/679 of May 4, 2016, On the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, art. 20, 2016 O.J. (L 119). The Regulation also includes several ancillary conditions, most notably that the required transfer is "technically feasible." *Id.* Further details can be found at *Guidelines on the Right to Data Portability*, DATA PROTECTION WORKING PARTY (Dec. 13, 2016), h ttp://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf; *see also*, Peter Swire & Yianni Lagos, *Why the Right to Data Portability Reduces Consumer Welfare: Antitrust and Privacy Critique*, 72 MARYLAND L. REV. 335 (2013).

that the anticompetitive dangers of lock-in outweigh the costs of sharing data across incompatible computing formats.[237] Suffice to say that publishing presents a particularly attractive case. We have argued that future customers are likely to invest at least as much effort in supplying data as publishers. From this standpoint, data portability is best understood as an attempt to restore the very conventional principle that those who invest in IP should also control it.

CONCLUSION: THREE FUTURES

Future technology advances will drive digital publishing in one of three directions. The first is well-established. Digital technology has helped consumers take over many search tasks formerly performed by publishers. At the same time, it has also eroded copyright and markups. This coincidence appears to have fortuitously improved society's IP tradeoff, although it remains possible that stronger copyrights – assuming they can be sustained against piracy – would do still better. Beyond this the rise of e-books offers a once-in-a-generation chance to dramatically expand readership through price discrimination. For now, the *Apple* decision stands in the way. But there is no evidence that the Second Circuit ever considered this issue, and the stakes are much too high to be decided by inattention. Judges urgently need to decide when and how firms should be allowed to negotiate contracts that jointly administer price discrimination across multiple markets.

In the longer run, Big Data methods will continue to improve. This will create ever-increasing pressures to simplify publishing's byzantine institutions. The question remains what should replace them. Our second future assumes that the gap will be filled in the conventional way by vertically integrated firms and joint ventures. Here, search would continue to be funded by IP, with occasional dilutions to entice additional data from readers. However, we can also imagine a third future in which Big Data develops in directions that replace large parts of today's ecosystem with volunteer-based open search collaborations. This will work best, and deliver the most benefits, to the extent that Congress follows up by narrowing copyright. To the extent this proves impossible, regulators should welcome private agreements to cut book prices for readers who donate data as an attractive second-best.

For now, judges and policymakers cannot anticipate, much less dictate, which of our three futures will prevail. Too little is known about how, or even whether, Big Data will fulfill its technological promise. Administrative agencies and courts can, however, set policies that let

[237] Swire & Lagos, *supra* note 236, at 337–38, 354−360. The cost would not necessarily be monetary. Portability could equally reduce firms' freedom to try new and better computing formats. *See id.* at 353−54.

better institutions emerge if and when they become possible. These measures should ideally include aggressive support for basic algorithm research; depositing government-supported research in the public domain; and clear-eyed antitrust policies that provide ample space for firms to share information. This leaves the much harder question of when companies that hoard data should be compelled to share. Americans should watch Europe's data portability experiment closely to see how well it solves this puzzle.