Yucesoy et al. *EPJ Data Science* (2018) 7:7
https://doi.org/10.1140/epjds/s13688-018-0135-y

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Success in books: a big data approach to bestsellers

Burcu Yucesoy[1], Xindi Wang[1], Junming Huang[1,2] and Albert-László Barabási[1,3,4,5*]

*Correspondence:
barabasi@gmail.com
[1]Center for Complex Network Research and Department of Physics, Northeastern University, Boston, USA
[2]CompleX Lab, Web Sciences Center, University of Electronic Science and Technology of China, Chengdu, China
Full list of author information is available at the end of the article

Abstract

Reading remains the preferred leisure activity for most individuals, continuing to offer a unique path to knowledge and learning. As such, books remain an important cultural product, consumed widely. Yet, while over 3 million books are published each year, very few are read widely and less than 500 make it to the New York Times bestseller lists. And once there, only a handful of authors can command the lists for more than a few weeks. Here we bring a big data approach to book success by investigating the properties and sales trajectories of bestsellers. We find that there are seasonal patterns to book sales with more books being sold during holidays, and even among bestsellers, fiction books sell more copies than nonfiction books. General fiction and biographies make the list more often than any other genre books, and the higher a book's initial place in the rankings, the longer the book stays on the list as well. Looking at patterns characterizing authors, we find that fiction writers are more productive than nonfiction writers, commonly achieving bestseller status with multiple books. Additionally, there is no gender disparity among bestselling fiction authors but nonfiction, most bestsellers are written by male authors. Finally we find that there is a universal pattern to book sales. Using this universality we introduce a statistical model to explain the time evolution of sales. This model not only reproduces the entire sales trajectory of a book but also predicts the total number of copies it will sell in its lifetime, based on its early sales numbers. The analysis of the bestseller characteristics and the discovery of the universal nature of sales patterns with its driving forces are crucial for our understanding of the book industry, and more generally, of how we as a society interact with cultural products.

Keywords: Success; Books; Bestsellers; Big data

1 Introduction

Books remain an important part of our lives, reading being the favorite leisure activity for many individuals. Indeed, the average American reads 12 to 13 books per year, and how people select the reading has been of much interest for researchers for decades. Consequently, book publishing is a huge industry in the U.S., with a revenue that is projected to reach nearly 44 billion U.S. dollars in 2020. In 2015, about 2.7 billion books were sold, a number that has remained fairly consistent in the last few years [1]. Of the over 3 million books in print in the U.S. every year, more than a hundred thousand are new titles. Yet, only a tiny fraction attract considerable readership. For example, less than 500 books make it to the New York Times bestseller lists and only a handful of authors stay on the

list for ten or more weeks. These near impossible odds reflect the challenges of capturing an audience in today's highly competitive world. Consequently, the success drivers of books remains of interest to many researchers [2]. Some of these drivers were explored in the literature [3, 4]: they are book critics [5], the author's and fans' circle of friends [6, 7], celebrities [8], online reviews [9, 10] and word of mouth [11]. The writing style of the author [12, 13], the amount of publicity [14], the timing of the book release [15], award-winning [16] or already bestseller [17] status of the book, the genre of the book and even the gender of the author [18, 19] are among the factors considered in past research. Yet, which books become successful and how they reach this status remains a mystery.

Aiming to address this mystery, here we explore the sales patterns of bestsellers and the authors who write them. We used a big data approach to understand the type of books that made it to the New York Times bestseller list as hardcovers, and quantified the sales patterns necessary to reach the lists and the characteristics of the bestselling authors. Additionally, we explore the weekly sale numbers [11, 20], allowing us to uncover the dynamic patterns of how a book becomes successful. This allows us to propose a statistical model that captures how collective interest in a book peaks and drops over time. The model accurately predicts the total number of copies an edition will sell, using the early sales numbers after the first release.

2 Data

The New York Times Bestseller List (NYTBL) is the most influential and prominent lists of best-selling books in the United States. Published since 1931, the list is digitally available since 2008 and consists of several sub-lists focusing on specific editions (hardcover, trade and mass-market paperback, e-book) and topics (fiction, nonfiction, children's and graphic books are the main weekly categories). For each book the list offers some basic identifying information like the ISBN number, title, author, publisher, amazon.com link. The NYTBL ranks books by the number of individual copies sold that week, using sales numbers reported by an undisclosed list of retailers across the United States, statistically weighted to represent all outlets nationwide [21]. The rankings are calculated each week, hence a book that continues to sell well stays on the same bestseller list for multiple weeks. In this study we consider all books featured on the New York Times *hardcover fiction* and *hardcover nonfiction* bestseller lists between August 6th, 2008 and March 10th, 2016 (410 weeks), altogether 2468 unique fiction and 2025 nonfiction titles.

To capture the time resolved sales patterns of books, we use NPD BookScan (formerly Nielsen BookScan), the largest sales data provider for the book publishing industry. Their database contains information about all print books that are being sold in the United States since 2004. For every book, this information includes the ISBN number, author name, title, category, BISAC number, publisher, price, total weekly sales across US and weekly sales in different geographical districts.

3 Bestselling books

Books come in a variety of formats and editions, from the more expensive hardcovers to the cheaper trade and mass market paperbacks, digital e-books and audiobooks. Despite the recent rise of digital books, printed books remain the preferred format for 65% of book readers in the U.S. Indeed, of the 2.7 billion books sold in 2015, 1.7 billion were printed books (577 million hardcover, 1.18 billion trade or mass market paperback). As new titles

are usually released in hardcover first, here we focus on the sales patterns of books that made it to the New York Times bestseller list as hardcovers in fiction and nonfiction categories, each encompassing a variety of genres. Not all of these genres are equally popular among the readers, some being more present on the bestseller list than others. Additionally, the dynamics of being featured on the list is different for each book: Some enter with low ranks and drop immediately off the list while others reach high ranks and stay on the list for a long time. In this section we focus on these dynamic patterns, along with the corresponding sale numbers and their seasonal patterns and we briefly discuss when and how different editions of the same title are released.

3.1 Genre fiction and memoirs dominate the bestseller list

According to a 2015 survey [1], mystery, thriller and crime are the preferred book genres in the U.S., nearly half of Americans reading in these genres. About 33% of the surveyed readers chose history as their favorite genre, while 31% preferring biographies and memoirs. To check if these preferences are reflected in the New York Times bestseller list, in Fig. 1(A) and (B) we break down all bestsellers by genre. We find that within fiction, most bestsellers fall into the 'general' fiction category (also known as 'mainstream' fiction), with 800 books making it the most popular category. This category mainly contains 'literary' fiction, i.e. fictional works focused more on themes and characters than on plot. These are the books frequently discussed by literary critics, featured in prominent venues and taught in schools, factors contributing to their popularity. In contrast 'genre' fiction, i.e. plot driven fiction like mystery or romance, shown separately in Fig. 1(A), is rarely considered by literary critics and is often reviewed only in venues catering to niche audiences. Yet, we find that the total number of bestsellers in these 'genre' categories collectively (1668) is more than twice the number of bestsellers in general fiction (800). Especially Suspense/Thrillers and Mystery/Detective categories resonate well with readers, in line with the survey findings [1]. Recent research and media discussions [18, 22, 23] noted that the popularity of genre fiction has been increasing over the years, thanks to the equal opportunities provided by online venues and rating systems, and the stagnant popularity of the traditional literary venues that remain focused on general fiction. We indeed observe a slight increase in the percentage of genre fiction among the bestsellers during the past decade (Fig. 1(C)).

Among nonfiction books, almost half of the 2025 bestsellers are from the Biography/Memoir category, consisting of books written by or about famous individuals, from politicians to artists or business personalities. Their dominance on the nonfiction market demonstrates a continuous interest in the life stories of well known individuals. The next most popular genre in nonfiction is history, in line with the survey results. The 'General Nonfiction' category encompasses multiple genres such as 'Psychology', 'Nature' or 'Philosophy' and 'Reference' mainly collects books in 'Science' or 'Technology'. For example, Malcolm Gladwell's bestselling books are categorized as General Nonfiction, while popular science books including Stephen Hawking's *The Grand Design* are classified in Reference.

In short, US readers prefer genre fiction over general fiction, making Thrillers and Mystery the most represented genres in the NYTBL over time. Biographies and memoirs are the most preferred genre within nonfiction, making up half of the nonfiction bestsellers.

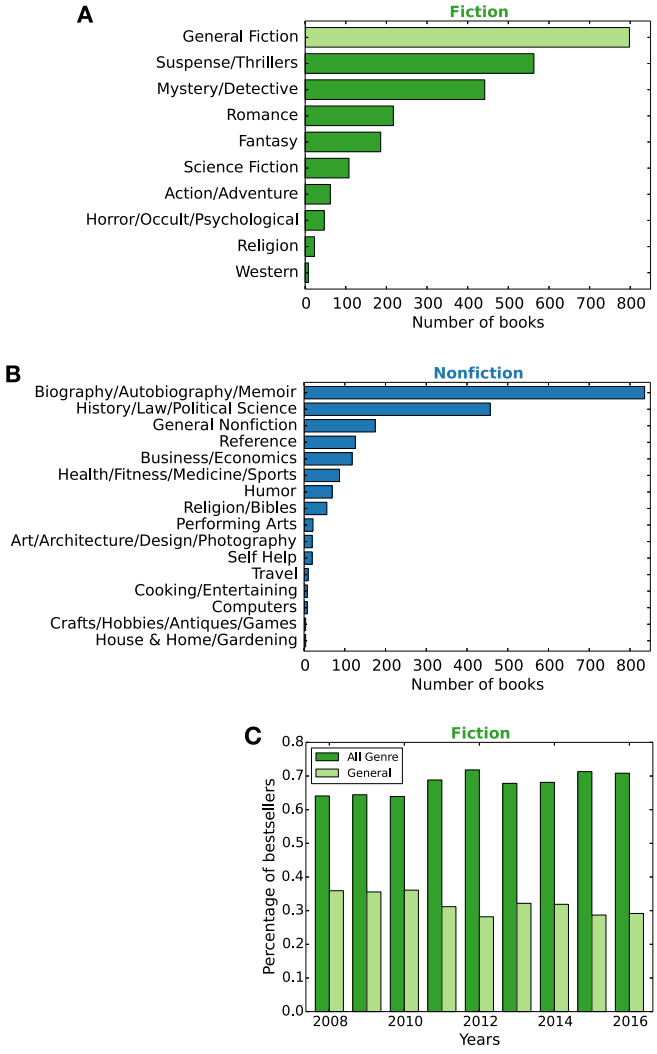

Figure 1 *Genre popularities*. The breakdown of bestsellers into genre categories of (**A**) fiction and (**B**) nonfiction. Most bestsellers fall into Suspense and Mystery genres combined in fiction or Biography/Memoir and History genres in nonfiction. (**C**) The percentage of genre fiction (all fiction that can be classified as a specific genre, shown in dark green) and general fiction (all else, shown in light green) on the bestseller list over the years. The popularity of genre fiction has been increasing steadily since 2011

3.2 Bestseller status rarely lasts

In marketing, a book is labeled a New York Times bestseller if it appears on the NYTBL for at least one week. Yet, there are major differences between bestsellers: some pop up on the list for a singe week while others retain their bestselling status for months and even years. To illustrate this, we measured the length of stay on the list for all New York Times bestsellers (Figs. 2(A) and (B)). We find that 25% of books appear only once on the list, while a few do spend an exceptional amount of time there.

For fiction, the number of books listed for only one week is high (26% of all books), indicating that the list changes considerably from week to week. In fact, only 10 of the 2468 fiction bestsellers stayed on the list for more than a year. The longest presence during our observation period is *The Help*, the 2009 book by Kathryn Stockett, which has been featured on the bestseller list for 131 subsequent weeks. Its continuous presence was helped

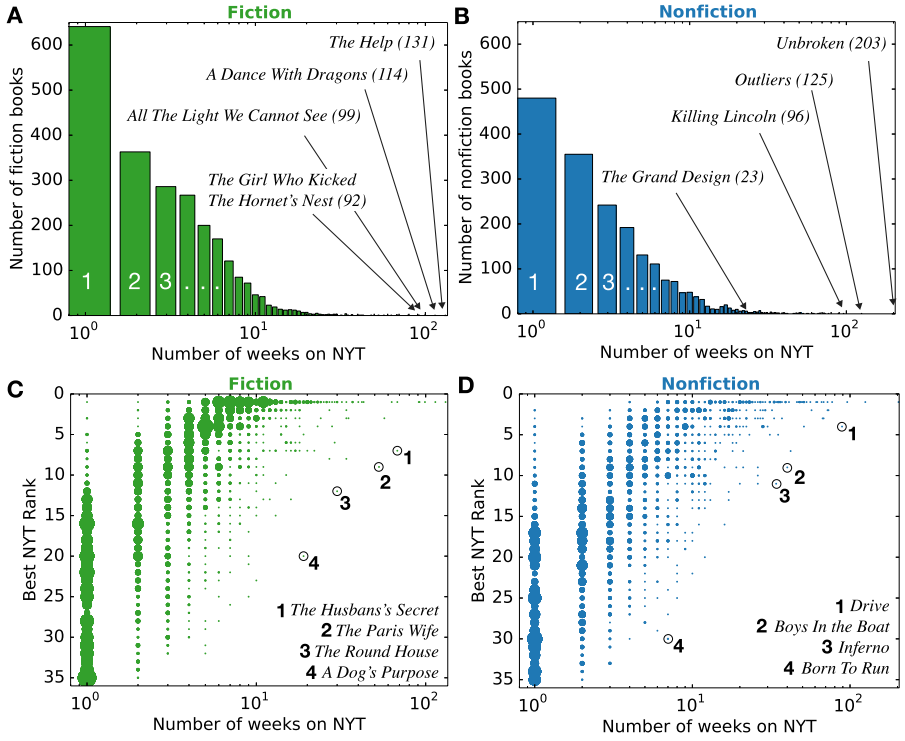

Figure 2 *Longevity of bestsellers*. Distribution of (**A**) fiction and (**B**) nonfiction bestsellers based on the number of weeks they stayed on the list. The number of weeks is shown in logarithmic scale to account for both the large number of short stays and the few exceptionally longer presence. We marked on the first columns the corresponding linear scale numbers for clarity. The best rank a book achieved on the New York Times (**C**) fiction and (**D**) nonfiction bestseller list vs. the number of weeks it stayed on the list. The size of the dots indicates the number of books with the same attributes. Overall, the better a book's best ranking, the higher is its probability of staying longer in the list

by a movie adaptation nominated for the Academy Award in 2011. Highly anticipated books in ongoing popular series tend to stay longer in the list, like the fifth book *A Dance With Dragons* in George R.R. Martin's *A Song of Ice and Fire* series with 114 consecutive weeks on the list, and the third book *The Girl Who Kicked the Hornet's Nest* in the *Millennium* series by Swedish writer Stieg Larsson with 92 weeks on the list. Finally, literary awards can also sustain bestseller status: *All the Lights We Cannot See* by Anthony Doerr, having won both the 2015 Pulitzer Prize for Fiction and the 2015 Andrew Carnegie Medal for Excellence in Fiction, was on the list for 99 weeks at the time of data collection.

In comparison, the nonfiction bestseller list shows slightly less variation from week to week, indicating that it is more common for nonfiction books to sustain their bestseller status. This is why we have fewer nonfiction books in our dataset than fiction books (2025 nonfiction bestsellers compared to 2468 in fiction). In the nonfiction category, 24% of books stayed only for one week on the list and 18 books lasted for more than a year. The most remarkable was *Unbroken: A World War II Story of Survival, Resilience, and Redemption* by Laura Hillenbrand which remained on the list for a record 203 weeks. Other examples of long-lasting success are *Outliers* by Malcolm Gladwell (125 weeks) and *Killing Lincoln* by Bill O'Reilly (96 weeks). In popular science category, *The Grand Design* by Stephen Hawking and Leonard Mlodinow stayed longest (23 weeks) on the NYTBL.

In general, the better a book's best ranking, the longer it stays on the NYTBL (Figs. 2(C) and (D)). The majority (86% for fiction and 89% for nonfiction) of the books that stayed on the list for a single week have reached the best ranking of 15 while the majority of the books (93% for fiction and 88% for nonfiction) that stayed for at least 10 weeks were ranked among the top ten at least once. Still there are a few exceptions, books that were ranked low on the list yet remained there for a long time. The most significant is *A Dog's Purpose* by Tom Doherty (now a motion picture as well) which remained on the list for 19 weeks even though its best position was number 20. *The Paris Wife* by Paula McLain or *Born To Run* by Christopher McDougall are also outliers that stayed much longer on the list than what would be expected based on their best ranking. Finally, the majority of the books that reached the top spot on the NYTBL stayed on the list for at least 10 weeks (51% for fiction and 80% for nonfiction).

In summary, most books stay on the NYTBL for only a week, and books lasting more than a year are extremely rare. That said, books reaching better ranks on the list stay on for longer periods compared to books ranked lower, many of the top ten books staying for several months at least.

3.3 Not all bestsellers sell

The number of copies a hardcover sells in its first year is an important measure of its commercial success. As after one year a cheaper paperback edition of the same title is likely to be released, the hardcover will no longer be the only print option. Therefore, in this section we focus on the first year sales of bestsellers, allowing us to explore their variability and the factors that determine their popularity. The one year sales distribution of all bestsellers indicates that the majority of bestsellers sell between 10,000 and 100,000 copies in their first year (Figs. 3(A) and (B)).

In fiction, *The Lost Symbol* by Dan Brown takes the lead with a record-breaking 3 million copies sold in a year, followed by the highly anticipated *The Girl Who Kicked the Hornet's Nest* and *Go Set a Watchmen*, selling over 1.6 Million copies each. These two books were anticipated for different reasons, the former being the third book in an ongoing successful series and the latter being the long-awaited second book of Harper Lee, published 55 years after her classic *To Kill A Mockingbird* (1960). There are also several books that even though made it to the NYTBL with high first week sales, could not sustain those numbers over the course of a year, such as *The Famous And The Dead*, being the conclusion to T. Jefferson Parker's *Charlie Hood* series (6 books) and *Hush Now, Don't You Cry*, the 11th novel in the *Molly Murphy Mysteries* by Rhys Bowen.

In nonfiction, the autobiography of former president George W. Bush, *Decision Points* sold the most copies in a single year, followed by the biography *Steve Jobs* by journalist Walter Isaacson, the basis for the 2015 movie of the same title. Yet, in the other extreme *The Slippery Year: A Meditation on Happily Ever After*, a memoir by Melanie Gideon sold less than 5000 copies in its first year. Occasionally nonfiction authors explore their themes throughout several books, resembling serialized novels of fiction, also resulting in high sales. A good example is *Killing Patton: The Strange Death of World War II's Most Audacious General*, by Bill O'Reilly and Martin Dugard about the final year of World War II and the death of General George Patton, which had sustained sales following other highly successful books with the same scheme by the same authors, *Killing Kennedy, Killing Lincoln* and *Killing Jesus*.

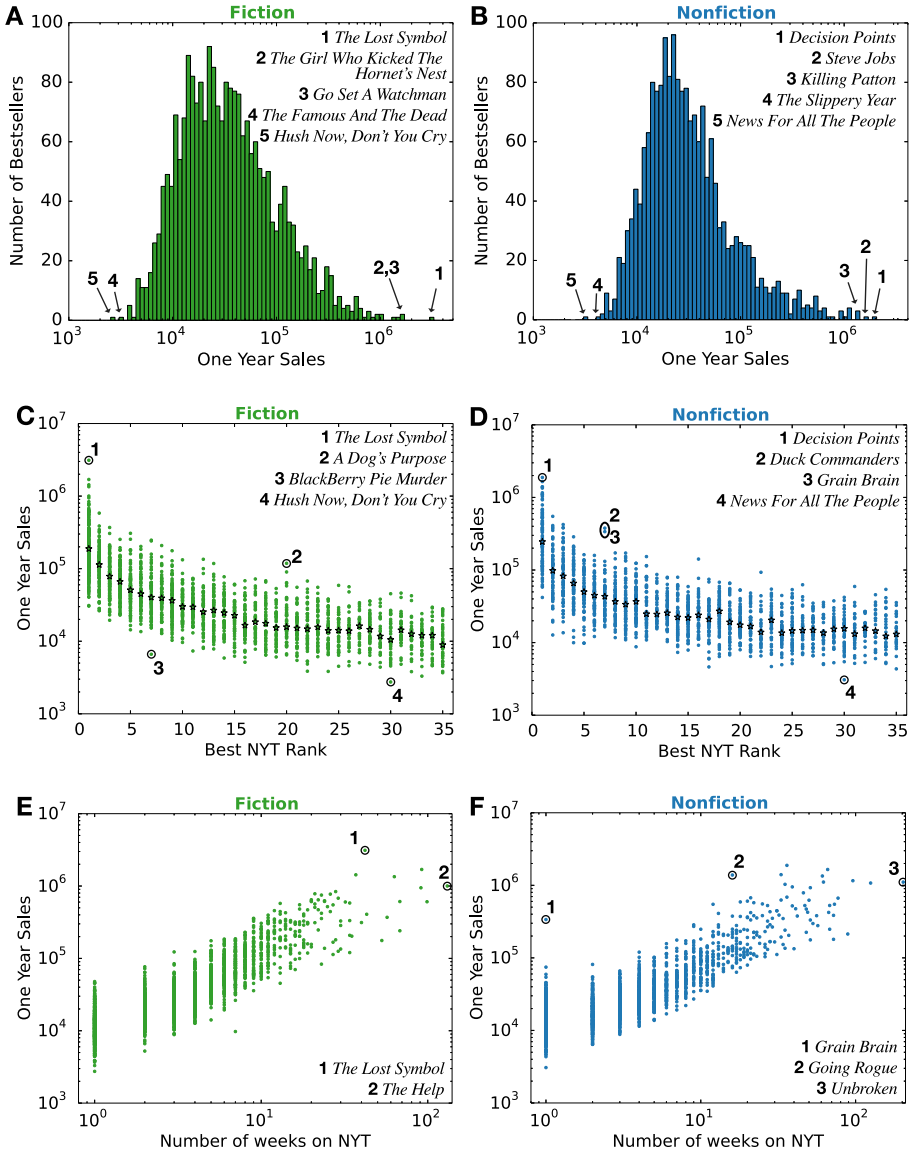

Figure 3 *Comparison of the NYTBL statistics and first year sales*. The distribution of the number of copies sold in a year for all hardcover bestsellers in (**A**) fiction and (**B**) nonfiction. The relation between one year sales and the best rank of the book in NYTBL (**C**) for fiction and (**D**) for nonfiction. The relation between one year sales and the length of stay on NYTBL (**E**) for fiction and (**F**) for nonfiction. Most hardcover bestsellers sell between 10,000 and 100,000 copies in their first year for both fiction and nonfiction, and the higher they get in the list and the longer they stay on, the more copies they sell in a year

To understand the dynamics of sustained sales, we looked into the relationship between the number of copies sold within a year and best ranking the book achieved on the list (Figs. 3(C) and (D)) and the length of stay of a book on the list (Figs. 3(E) and (F)). Obviously, the more copies a book sells in a single week, the better is its ranking in the bestseller list. For most books we also observe a direct correlation between the best ranking and the number of copies sold within a year. The most remarkable are *The Lost Symbol* and *Decision Points*, books ranked number one on the NYTBL in their category (fiction and nonfiction respectively) selling more than a million copies in their first year after pub-

lication. Yet we do note some outliers, selling significantly more (*A Dog's Purpose* in fiction and *Duck Commanders* and *Grain Brain* in nonfiction) or significantly less (*Blackberry Pie Murder* and *Hush Now, Don't You Cry* in fiction and *News For All The People* in nonfiction) copies in a year than would be expected from their best positions in the NYTBL. In case of the *Duck Commanders*, a behind-the-scenes account of the family featured in the reality TV show Duck Dynasty, the sustained sales were supported by the TV show's continuous popularity, prompting several books from different members of the same family with various degrees of success. *Grain Brain* by neurologist David Perlmutter is also an interesting case showcasing the seasonality of the bestseller lists. The book first came out in September 2013 and hit the NYTBL soon after, reaching its highest sales in December when book sales are typically the highest (see the following section). Hence, despite the impressive sales numbers, in those weeks it did not qualify for better rankings in the NYTBL, even dropping entirely from the list.

Not surprisingly, we also observe a direct correlation between the length of stay on the bestseller list and the number of copies sold in a year (Figs. 3(E) and (F)). The only two exceptional cases are in the nonfiction category. As already mentioned, *Grain Brain* had sustained high sales but a short stay on the NYTBL due to the seasonality of the list. In case of Sarah Palin's *Going Rogue*, the book sold a record-breaking 500,000 copies in a single week, but the sale numbers dropped steadily afterwards. Yet the sales accumulated in the 16 weeks the book was on the list were sufficiently high to make the book an outlier among bestsellers.

In summary, the number of copies a NYTBL book sells in its first year spans over two orders of magnitude. Overall, the best rank a book achieves on the list is a good predictor of its yearly sales: the better the rank, the higher the total sales. The length of stay of a book on the list is another good indicator of the number of copies sold, as longer stays mean larger number of copies sold every week.

3.4 We read more fiction than nonfiction and buy books during holidays

Since 2008, books on the hardcover NYTBL have sold anywhere between a thousand to a million copies in a single week. We show the distribution of the weekly sales that have gotten these books to the list in Fig. 4(A) and (B). Of course a book ranked first must sell far more copies than a book ranked 20 or 35. Accordingly, the high end outliers in Fig. 4(A) and (B) were all ranked number one for several weeks. The low outliers are included on the list either by mistake, or BookScan has a different record of their sales from New York Times. They are all books with much higher sales on other weeks, but on those particular weeks, their sale numbers recorded by BookScan were much lower than what is typically needed to hit the NYTBL. Aside from the extremes and differences between ranks, there may be several causes for the general high variability, as we discuss next.

First, we looked into the sales needed to make it to the bestseller list since 2008. In Fig. 4(C) we consider week 33 (August), showing the number of copies each book on the NYTBL sold that week each year. We see that fiction books sell more copies than nonfiction books, in other words, fewer copies are needed to qualify a book for the nonfiction list than the fiction one. Also the stability of the year-by-year sales pattern is remarkable: today a book needs to sell between a 1000 to 10,000 copies to make it to the bestseller list, a range that stayed roughly the same during the past eight years.

Next, we looked into the seasonal fluctuations in the sales patterns during a year. To explore how these fluctuations affect the bestseller list, we measured the median number

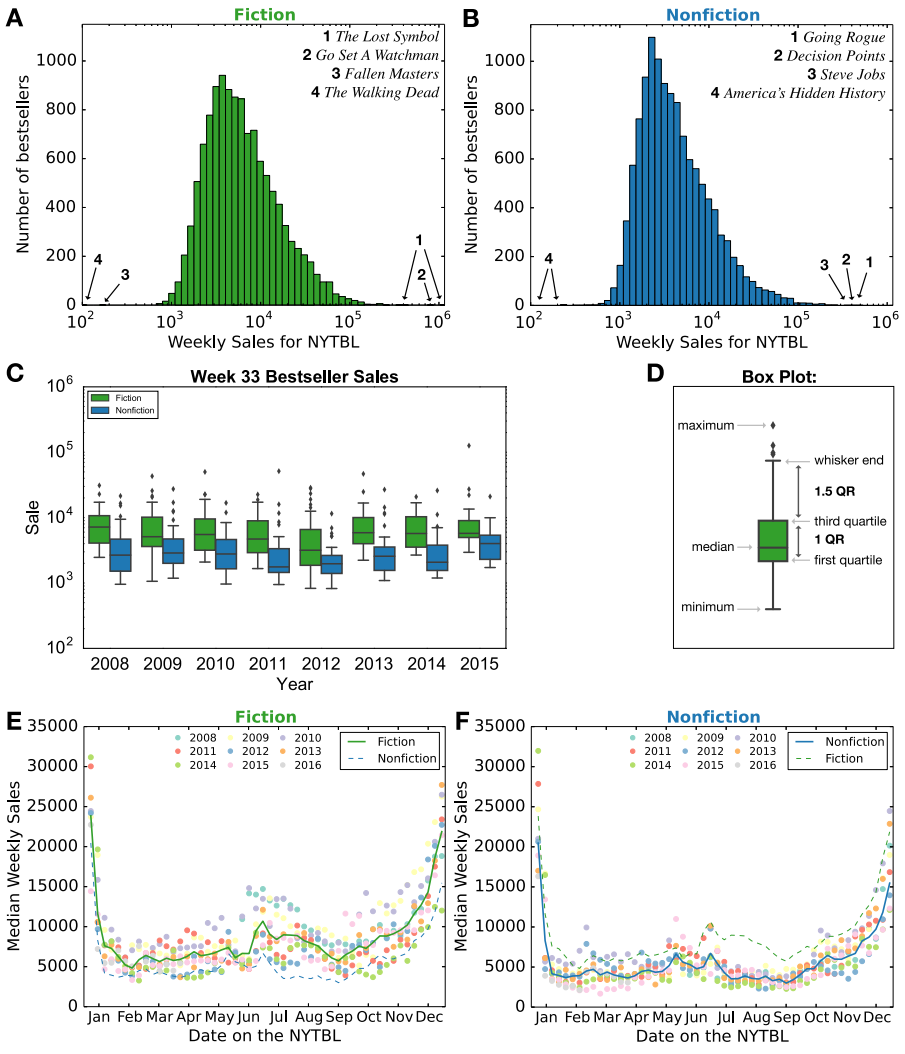

Figure 4 *Making the NYTBL*. The distribution of the number of copies sold in a week needed to be featured in NYTBL for (**A**) fiction and (**B**) nonfiction. (**C**) The number of copies bestselling books have sold in a single week in mid August (week number 33) that got them to the list over different years. (**D**) Explanation of the box plot technique used in (**C**). The median number of copies the bestselling books sold to hit the NYTBL during different weeks throughout a year, for (**E**) fiction and (**F**) nonfiction. In general, books need to sell between 1000 and 100,000 copies in a single week to hit the NYTBL, a range that has been fairly stable over the years. Fiction book sales are higher in summer and all sale numbers are significantly elevated from December to early January. Fiction sells more than nonfiction throughout the year but the gap is smallest in early January

of sales that got the books on the list at different times of the year (Figs. 4(E) and (F)). The dots correspond to the median sales of all books on the NYTBL at any given week each year, the line indicating the median over all years. We focus on the median instead of the average given the high variability amplified by record-breaking sales highlighted in Figs. 4(A) and (B).

Overall, we find that median sales mostly fluctuate between 4000–8000 in fiction and 2000–6000 in nonfiction. Yet, there is a significant increase in sales late-December during holiday shopping, a pattern persisting into early January, likely due to delays in sales reporting. In early January, the lowest median sales over the years is close to 15,000 copies a week, a number higher than the highest median sales of any other time of the year ex-

cept late December. For fiction, a similar but less pronounced peak is observed during the summer months with median sales surpassing 10,000, likely due to book purchases in preparation for the summer vacation. In nonfiction, there is no such summer peak. During these periods of elevated sales a book needs to sell more copies to make it to the New York Times bestseller list than during other months. We also note that in general, fiction books sell more copies than nonfiction, a gap which is largest during summer and decreases considerably during the holiday season, where the sales of both fiction and nonfiction are significantly elevated.

In summary, we find that books featured in the NYTBL over the years hit the list by selling anywhere between thousands and tens of thousands copies, a range that has been stable since 2008. Seasonal fluctuations within a year matter much more, books needing higher sales during the holidays to stand out, even though more books are purchased in that period. Additionally, a book on the fiction list needs to sell more copies on average compared to the nonfiction bestseller list, due to the fact that on average, fiction sales are higher than nonfiction sales.

4 Bestselling authors

It is sufficient for an author to have written a single book that appeared on the NYTBL for a single week to be labeled a 'bestselling author', a label that sticks for life. Yet, not all bestselling authors are alike. There are those with a single high selling book in their career, like Kathryn Stockett (*The Help*), and there are authors with over fifty books with varying sale numbers under their belt, like James Patterson or Stephen King. Additionally, some authors build their readership over time, achieving bestseller status with their later work while others enter the NYTBL with their first book. The success of a book is deeply linked to the previous success and the name recognition of its author, prompting us to explore in this section the dynamics of success for authors, quantifying the differences between authors in terms of productivity, repeat success and gender among authors within different bestseller categories.

4.1 Fiction authors have more repeat success than nonfiction authors

To understand the patterns of productivity among bestselling authors, we collected all unique titles published by them in hardcover since 2008, regardless of whether they made the bestseller list or not. After eliminating new editions of older titles, we ended up with 5396 books (2468 of them bestsellers) by 854 authors with bestsellers in fiction and 3968 books (2025 of them bestsellers) by 1670 authors with bestsellers in nonfiction categories. These numbers already indicate that fiction authors are more prolific than nonfiction ones, with half the number of fiction authors having written 1.5 times more books than nonfiction authors since 2008.

As indicated by Figs. 5(A) and (B), only 14% of the fiction authors have written only one book since 2008. The vast majority of them have at least two books, but having close to 10 books is also common. Some authors are significantly more productive than others, with James Patterson being an outlier: he published 94 hardcovers since 2008, often with coauthors and in a variety of genres such as mystery, suspense, romance and even nonfiction. In Fig. 5(C) we take a closer look at his productivity and sales patterns, denoting each hardcover with a line located at its publishing date, its height indicating the number of copies sold in its first year. His most successful book was *I, Alex Cross*, the 16th novel in his *Alex*

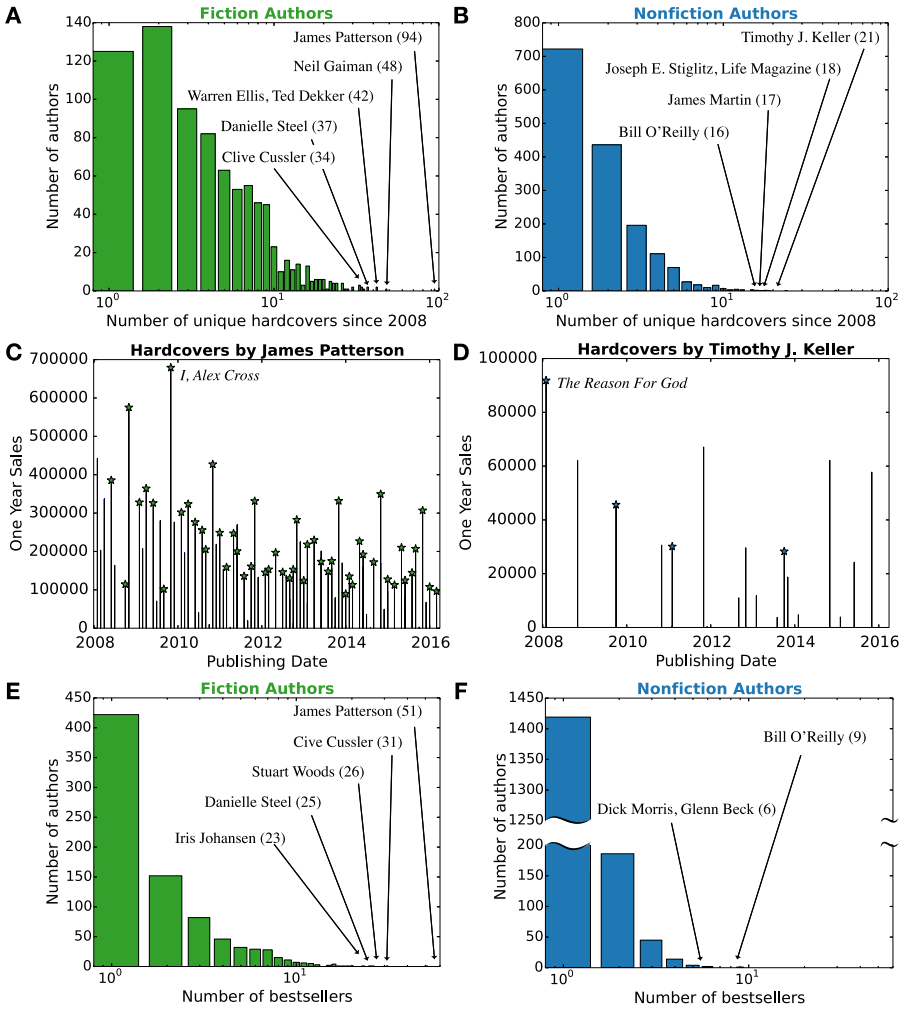

Figure 5 *Productivity and repeat success*. The number of unique hardcovers published by (**A**) fiction and (**B**) nonfiction authors since 2008, including their bestsellers. Books published by (**C**) James Patterson and (**D**) Timothy J. Keller, the line height indicating one year sales and stars marking bestseller status. The number of (**E**) fiction and (**F**) nonfiction bestsellers per author. While most authors in both fiction and nonfiction have only one bestseller, repeat success is more common for fiction authors than for nonfiction ones. Additionally, having one or two hardcovers published since 2008 is the norm for most authors, yet bestselling fiction authors publish significantly more compared to nonfiction authors

Cross series published in 2009, which stayed on the NYTBL for more than 20 weeks. With stars indicating bestseller status, we see that more than half (51) of his books were bestsellers. Fiction authors who also write graphic novels, like Neil Gaiman (48 books) and Warren Ellis (42), are also productive due to the usual high volume of publications in the graphic novel category. Other exceptionally prolific authors are mystery, thriller and fantasy author Ted Dekker (42), romance author Danielle Steel (37) and thriller author Clive Cussler (34).

In nonfiction, high productivity is rare (Fig. 5(B)) with nearly half (43%) of the authors having published only one hardcover since 2008. The most prolific author in nonfiction is pastor and theologian Timothy Keller, having written 21 books about spiritual topics. In Fig. 5(D) we show his career since 2008. He had 4 bestsellers, starting with the 2008

book *The Reason For God*, with the most successful first year sales. Even though several more of his books sold quite well over the course of a year, their individual weekly sales were not sufficiently high for them to make the NYTBL in any particular week. Economist Joseph E. Stiglitz is the next most productive nonfiction author with 18 books, including several textbooks. The editors of Life Magazine have curated 18 books since 2008, primarily focusing on events and people of public interest, like the sinking of Titanic or the life of Barack Obama. Five of them became bestsellers in nonfiction category.

Next, we look at the repeat success by only considering bestsellers (Figs. 5(E) and (F)). In fiction, we find that the 2468 hardcovers on the NYTBL are written by only 854 authors, indicating that the list is dominated by a small number of authors with multiple bestsellers. We have already seen that James Patterson takes the lead in both repeat success and productivity. Similarly, Clive Cussler with 31 bestsellers and Danielle Steel with 25 bestsellers are also especially successful, in addition to being rather prolific. In general, bestselling authors with multiple books will often have multiple bestsellers. This is partly because fiction authors commonly write novels in serialized form and once a series builds up an audience, the subsequent books in the series also receive substantial attention.

In nonfiction NYTBL, repeated authorship is less common: the 2025 nonfiction books are written by 1670 authors, indicating fewer recurrent authors. The distribution in Fig. 5(F) indicates that an overwhelming majority (85%) of the bestselling nonfiction authors have only one bestseller since 2008. Interestingly, in nonfiction, the top 3 authors with most bestsellers, Bill O'Reilly, Dick Morris and Glenn Beck, are all political commentators, writing mostly about current public affairs and government issues in the U.S. Yet, as we mentioned earlier, Bill O'Reilly has also written several bestselling books with historical themes, speculating about the deaths of prominent historic figures, with co-author Martin Dugart.

In summary, fiction authors are likely to write multiple books in quick succession and often in serialized format, and they often have multiple bestsellers. In nonfiction however, the norm is one bestseller per author, which is typically the only hardcover they published since 2008. This is partly due to the fact that most nonfiction bestsellers are memoirs, books written by or about famous individuals, without repeat authorship. Yet higher productivity and repeat success does happen in nonfiction as well, albeit less frequently.

4.2 Many bestsellers are debut books but late success is also possible

To understand the patterns of writing a bestselling book, we need to explore the careers of individual authors. In this section we ask if it is more common for authors to reach bestseller status with their first book, or if bestselling status is something built up through the increasing popularity of multiple books. To address this question, we focused on bestselling authors who started publishing in 2008 and after, a cohort of 145 authors in fiction and 591 authors in nonfiction.

In fiction, the author with the most hardcovers is Taylor Anderson, with 10 books in *The Destroyermen* series (Fig. 6(A)). The first three books in the series sold relatively poorly, yet we observe increasing sales with each new book. It was the 4th book that reached a mass audience, doubling the weekly sales at its release, yet still not enough to land on the bestseller list. The fifth book of the series finally became a bestseller, yet for a single week. Yet, the added visibility propelled the 6th and 7th books in the series to the NYTBL. His following 3 books, although selling strongly, did not make the list any longer. Another

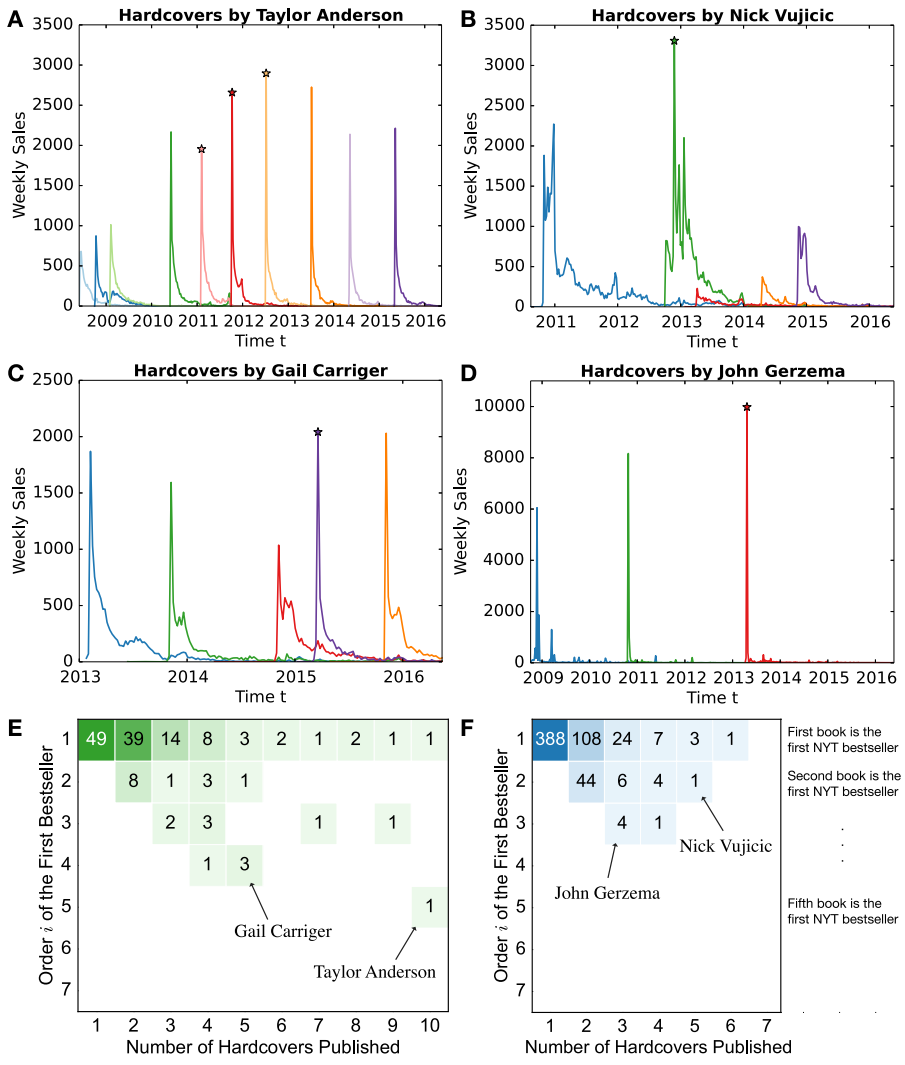

Figure 6 *When do authors succeed?* (**A**) The sales history of *The Destroyermen* series author Taylor Anderson, showing for each book the weekly hardcover sales. Curves marked with a star on top indicate bestsellers. Same for authors (**B**) Nick Vujicic, (**C**) Gail Carriger and (**D**) John Gerzema. (**E**) Heatmap showing the number of hardcovers a fiction author has published in their career vs the order i of one that became the first bestseller, (**F**) same for nonfiction authors. Only authors who started their career in 2008 and after are considered and boxes the authors shown in (**A**)–(**D**) belong to are marked accordingly. We see that while most bestselling authors first get on the NYTBL with their first book, later success is not uncommon

fiction author achieving success with late books is the author of the young-adult steampunk *Finishing School* series Gail Carriger (Fig. 6(C)). She already had moderate success in paperback form with her adult oriented *Parasol Protectorate* series, yet none of her four hard cover format *Finishing School* novels (shown in blue, green, red and orange) made the NYTBL. Her 2015 hardcover *Prudence* (purple) was the book that finally landed her on the list.

In nonfiction, Christian evangelist and motivational speaker born with tetra-amelia syndrome (a rare disorder characterized by the absence of arms and legs) Nick Vujicic is one of the more productive authors starting his writing career in 2010 (Fig. 6(B)). His first book, *Life Without Limits*, was an international success, being translated into more than

20 languages. Yet, it did not make the NYTBL. His second book, *Unstoppable*, got there two years later, helped by the buzz created by his first book and possibly his motivational speaking engagements. He went on to write three more books, none of them matching the success of his first two. Another nonfiction author, columnist and businessman John Gerzema writing about impact of leadership ethics, had his success grow steadily with each subsequent book, finally getting the third one, *The Athena Doctrine*, into the NYTBL (Fig. 6(D)).

To understand the typical order of a bestseller within an author's career, we constructed heatmaps showing the distribution of the place (index) of an author's first bestseller among their published hardcovers with respect to how many hardcovers they published in total (Figs. 6(E) and (F)). We see that many (corresponding to the sum of the first row in the Figs. 6(E) and (F), 120 for fiction and 531 for nonfiction) of the bestselling authors who debuted on or after 2008 first got on the NYTBL with their first book. This is partly due to the fact that debut novels are over-represented in this selection, since we did not consider authors with previous publishing history even if their first bestseller was published after 2008. Consequently, we observe that many of the considered bestselling authors had only one book so far, overwhelmingly so for nonfiction authors (65%). Yet, late success like Taylor Anderson's is not unheard of. Many 2-book authors got into the NYTBL with their second book (8 in fiction and 44 in nonfiction) and even later success is achieved by several (12 in fiction and 5 in nonfiction). Additionally, there are 2 fiction authors like Gail Carriger with 5 books whose 4th book was their first bestseller and 3 nonfiction authors like John Gerzema who built up to bestselling status with the first two of their three books.

In general, we find that most bestselling authors who started their careers on or after 2008 were successful with their first book, yet getting into the NYTBL with a second or later book is possible as well.

4.3 Previous success is a good predictor of the next book's success

As we established earlier, repeat authorship is common on the bestseller list, particularly in fiction. We have also seen that some authors start with a large readership while others build their readership gradually, and finally some lose the public's interest after a few successful books. To quantify how the sales of a previous hardcover affect the sales of the subsequent book, in Figs. 7(A) and (B) we show the one-year-sales of hardcover number i vs. one-year-sales of the previous hardcover, $i - 1$. Since we are showing only unique hardcovers from bestselling authors after 2008, it is not surprising that most books are concentrated on the top right of the plot indicating high sales, a pattern present in both fiction and nonfiction. The fact that most books fall very close to the 45 degree line indicates that books that sell well are likely to be followed by books with comparably strong sales. Consequently, we find that best selling authors tend to sustain their success. Yet, even bestselling authors can publish books that sell poorly, as indicated by the outliers marked in Figs. 7(A) and (B), showing no clear relation between a book's and its predecessor's sales.

In general, the outliers are books in other genres than the one the author had their bestseller in. Remarkable examples are *The Expo Files*, the only nonfiction book by the *Millennium Triology* author Stieg Larsson (marked 1 in Fig. 7(A)), children's books by *The Truth According to Us* author Annie Barrows (2) or by John Jackson Miller (4), author of several *Star Wars* novels. We see the same trends in Fig. 7(B) for nonfiction, with a song book by Kristin Chenoweth (1) whose memoir was a bestseller, children's books by Kay

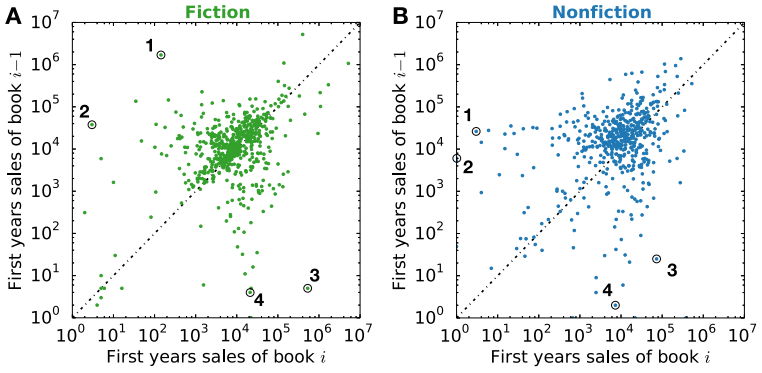

Figure 7 *Previous sales matter*. The relation between the sales of the previous hardcover $i-1$ and the sales of the next one, i, by the same author, for (**A**) fiction and (**B**) nonfiction. Best selling authors tend to sustain their success with new books, as most dots lie close to the 45 degree line. The outlier books, discussed in the main text, are written by (1) Stieg Larsson, (2) Annie Barrows, (3) Melissa de la Cruz and (4) John Jackson Miller for (**A**) and (1) Kristin Chenoweth, (2) Kay Robertson, (3) Eric Blehm and (4) Lady Colin Campbell for (**B**). They are either books written in a different genre from the previous bestselling books of the author, or reissues of old books with a different title

Robertson (2) of the Duck Dynasty fame or by Eric Blehm (3) who had multiple nonfiction bestsellers about life in the military or an autobiography from the Lady Colin Campbell (4) who normally writes about the British Royal Family. Finally, reissues of old books with new titles do not sell well regardless of an otherwise successful author, as shown by the Melissa de la Cruz book (3 in Fig. 7(A)) *Popularity Takeover*, originally published as *Lip Gloss Jungle* 6 years prior.

In short, bestselling authors are likely to sustain their success with subsequent books unless they choose to write books in significantly different genres, losing their reader base.

4.4 Female authors dominate romance while men dominate nonfiction

Books are cultural products and as such, their success is highly interconnected with the cultural makeup of our society. Since gender and gender role perceptions are an essential part of this makeup, we next explore if gender plays a role in book success.

We start by dividing the books into four categories: books written by male authors (2182 books), by female authors (1518), books with multiple authors (multi, 720) and finally books for which we could not identify the author's gender (unknown, 75). Multiple authorship means either that several authors collaborated on a book (most books by celebrities fall into that category, being co-written with professional writers), or they may be translations to English, hence often the translator is listed as a co-author. In rare cases a book written under a pseudonym may have two names listed once the pseudonym is unveiled, like *The Cuckoo's Calling* written by J.K. Rowling under the pen name Robert Galbraith. To obtain gender information from author names, we used a database of first names separating names into groups of 'male', 'female' and 'neutral', and verified a selection from author informations available on Wikipedia and GoodReads.

In fiction, we find that bestselling books are mostly written by a single author, and the number of bestsellers written by female authors and male authors are indistinguishable (Fig. 8(A)). In nonfiction, however, the bestseller list is dominated by male authors (Fig. 8(B)). As we do not know the ratio of female/male authors publishing in nonfiction that do not make the NYTBL, we cannot tell if readers prefer nonfiction written by men, or

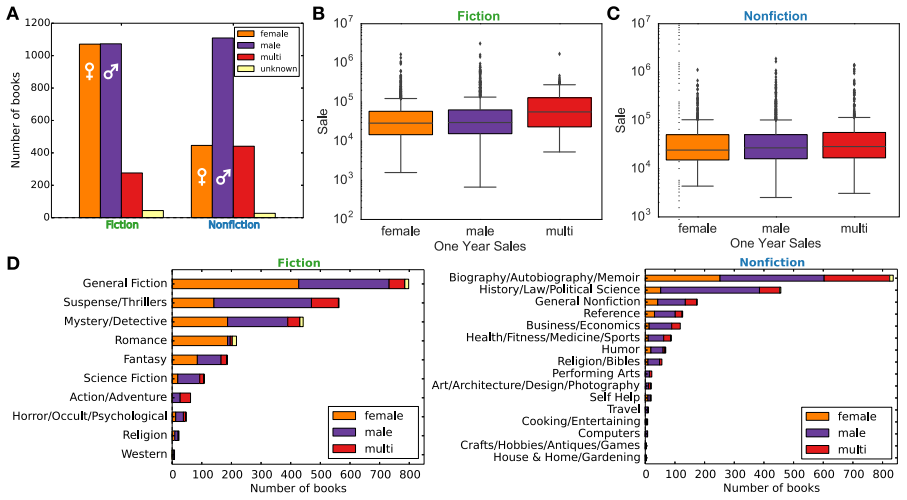

Figure 8 *Bestsellers and gender.* (**A**) The breakdown of the New York Times hardcover bestseller list by the author's gender. Fiction bestsellers are equally likely to be written by female or male authors, while most nonfiction bestsellers are written by male authors. (**B**) One year sales of the fiction bestsellers written by different gender groups, (**C**) the same for nonfiction. A breakdown of the bestselling book categories by gender, fiction (**D**) and nonfiction (**E**). Median yearly sales for books written by different groups are indistinguishable. In fiction, romance genre is dominated by female authors, while male authors are more prominent in suspense/thrillers, and in nonfiction the Biography/Memoir genre has the most female presence

simply more nonfiction is written by men. If we compare one-year-sales of the bestsellers written by different groups, we do not see significant gender differences, as indicated by the median values shown as stars in Figs. 8(C) and (D). In other words, in both fiction and nonfiction, the sales patterns of female and male authors are largely indistinguishable.

To observe if different gender authors prefer different genres, we broke the bestselling book categories by gender (Fig. 8(D)). We find that female authors are better represented in Literary (General) Fiction than men and dominate in Romance. In contrast, Thrillers, Science Fiction or Action/Adventure are dominated by male authors. Both groups are represented equally in Mystery novels. In nonfiction, all categories are dominated by men (Fig. 8(E)), but for Memoirs, the gender difference is smaller.

In summary, we find that in fiction, equal number of bestsellers are written by female and male authors, and bestsellers with multiple authors are rare. In contrast, male authors dominate the nonfiction NYTBL, and the number of female- and multi-authored bestsellers are similar to each other.

5 The dynamics of book sales

In this section we model the temporal changes in book sales, allowing us to capture and predict the observed sales patterns.

5.1 Bestsellers reach their sales peak in less than ten weeks

In Sect. 3 we argued that the first year sales are the most important for a hardcover. Indeed, for the 2035 fiction bestsellers we have at least two years of sales data, we find that 96% of the sales took place in the first year. Similarly, 94% of the sales of 1699 nonfiction bestsellers also happen in the first year. To systematically explore the dynamics of the sales patterns, we start by showing the weekly sales of all bestselling fiction (Fig. 9(A)) and nonfiction

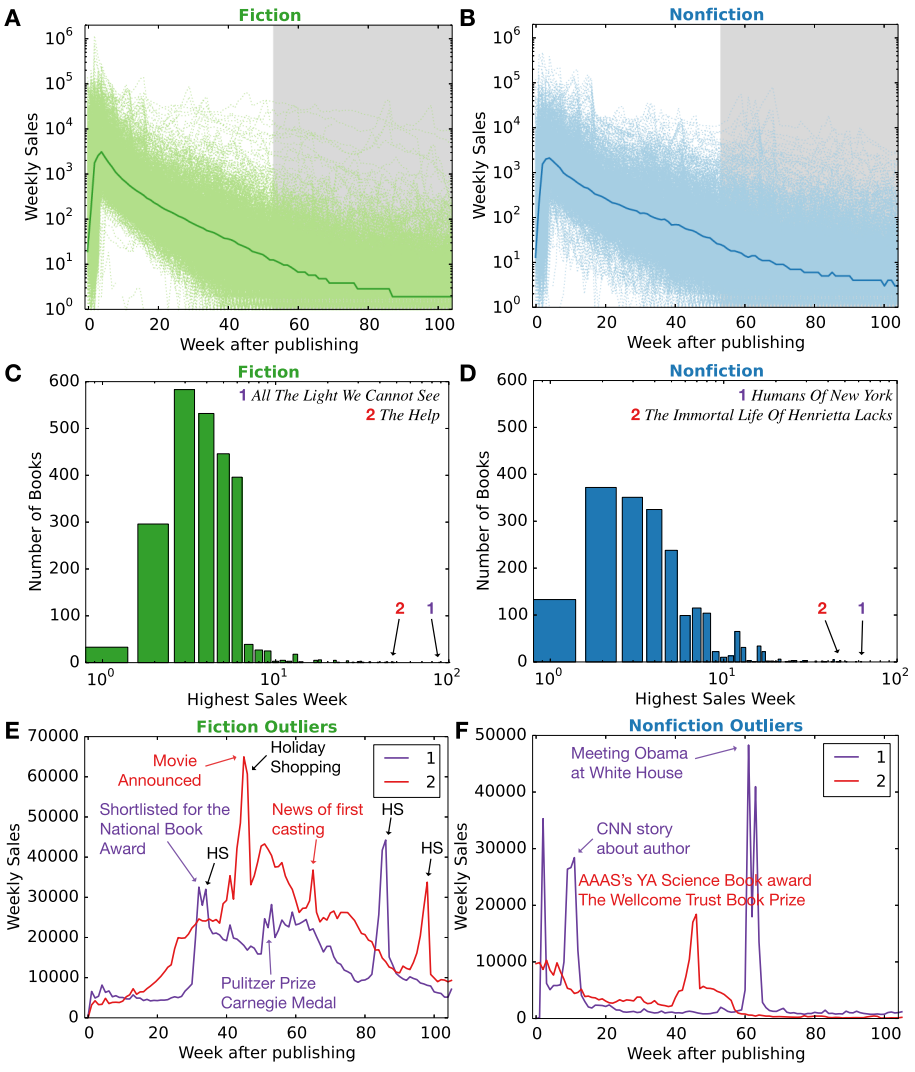

Figure 9 *Weekly sales*. The weekly sales numbers of all bestselling (**A**) fiction and (**B**) nonfiction books, thick lines showing the median sales values for that week. The shaded area starts after the first year and only 4% of fiction and 6% of nonfiction sales happen there. The number of books reaching their peak sales at a given week after publication, (**C**) for fiction and (**D**) for nonfiction. While the sales numbers vary significantly from book to book and from week to week, bestseller sales in both fiction and nonfiction increase sharply at first, reach their peak in the first ten weeks and drop slowly afterwards. The sale curves of selected outliers reaching their peak sales later than usual are shown in (**E**) for fiction and (**F**) for nonfiction

(Fig. 9(A)) books. The thick line corresponds to the median sale values. The peak sale values vary significantly from book to book, some books selling over 100,000 copies at their peak while others only reaching a few hundreds. We therefore use a logarithmic scale to display all sales curves. These plots already indicate that for both fiction and nonfiction peak sales are within the first ten weeks after a book's release.

We find that almost all books, regardless of category, peak in the first 15 weeks after publication (Figs. 9(C) and (D)). Furthermore, most fiction books have their peaks strictly in the first 2–6 weeks; in contrast for nonfiction, even though peaks at weeks 2–5 are common, the peak can happen any time during the first 15 weeks. For example both *The Lost Symbol* by Dan Brown and *Go Set A Watchman* by Harper Lee peaked in their 3rd

week, and so did Sarah Palin's *Going Rogue*. George W. Bush's *Decision Points* had its peak sales even earlier, on the second week after publication.

Still there are some outliers that peak much later, towards the end of their first, or well into their second year. These exceptionally late peaks are typically triggered by exogenous events such as winning awards, being adapted for a movie or in rare cases, having a prominent public figure's endorsement. We are showcasing several such examples in Figs. 9(E) and (F). *All the Light We Cannot See*, shown in purple in Fig. 9(F), is a novel written by Anthony Doerr, published on May 6, 2014. The novel had an initial peak and subsequent decline when it was shortlisted for the National Book Award later that year. The sales numbers tripled the week after it lost the award to *Redeployment* by Phil Klay. The novel later won both the 2015 Pulitzer Prize for Fiction and the 2015 Andrew Carnegie Medal for Excellence in Fiction, causing further peaks in sales numbers, but the most drastic effect was seen at the end of 2015, where people overwhelmingly chose this multiple award-winning book during their holiday shopping. Another example of awards causing late peaks is the nonfiction book *The Immortal Life Of Henrietta Lacks* (red in Fig. 9(F)) by Rebecca Skloot, which won both the American Association for the Advancement of Science's Young Adult Science Book award and the Wellcome Trust Book Prize. *The Help* (red in Fig. 9(E)) by Kathryn Stockett on the other hand was a 'sleeper hit' which gradually increased in sales until a movie adaptation was announced. The announcement, coinciding with the holiday shopping season, propelled the book's sales to more than 60,000 a week. Another peak in sales happened when the first pictures of the movie's cast appeared and the following holiday shopping season was also beneficial for the book. Finally, *Humans of New York* author Brandon Stanton (purple in Fig. 9(F)) and his well-known Facebook page of the same title as the book were featured on CNN shortly after the book's publication, causing a second peak in sales. But the book's biggest success came when Stanton interviewed the then U.S. President Barack Obama in the Oval Office in January of 2015.

These exogenous events aside, the data indicates that the first few weeks of a book are crucial: This is when the books capture the interest of their readership. Also this is the time when publishers will invest in a book's advertising and the most likely period for a book to be featured in the front of book stores and considered for reviews in various media. As such, a book's sales to be the highest in that period.

5.2 Sales follow a universal pattern

As can be seen in Figs. 9(A) and (B), most books follow a similar sales pattern: the sales increase very fast, reach their peak in the first ten weeks and drop dramatically afterwards. This similarity suggests the existence of a universal sales pattern, i.e. the possibility that the properties for all sales curves are the same, independent from the details and degree of complexity of each individual book's sales narrative. This hypothesis allows us to develop a simple yet general model, helping us identify the mechanisms that drive the sales of books. In general, three fundamental mechanisms contribute to the observed sale patterns:

(i) Each book carries a different value for its audience, stemming from the author's name recognition, the writing style, the marketing efforts by the publisher and even the quality of the book cover. Some books are anticipated and well-liked, resulting in high sales, some will be unexpected, lacking familiar elements and hard to get into, resulting in lower sales. To account for these inherent differences, we define a parameter called the book's *fitness*, η_i, that captures the book's ability to respond to the taste of a wide readership.

(ii) Second, a book that sells well will attract even more sales, an effect called *preferential attachment* [24, 25]. Preferential attachment in this context is likely rooted in collective effects, like recommendations from friends, critics, celebrities, online reviews and bookstores who display a sought after book in visible spots. Mathematically it implies that the likelihood of purchasing a book depends on its up-to-date sales, S_i^t.

(iii) Finally, even the best books lose their novelty and fade from the public eye some time after their publication. Barring exogenous events, once the book reached its target audience, less and less individuals are interested in purchasing it. To model this gradual loss of interest, we need to add an *aging* term, using a form adopted from the decay of citations in research papers [26]

$$A_i(t) = \frac{1}{\sqrt{2\pi}\sigma_i t} \exp\left[-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}\right], \tag{1}$$

where μ_i is the book's immediacy determined by the time the sales reach its peak and σ_i is the decay rate capturing the longevity. In case of books, a lognormal aging term (1) is motivated by the fact that the time of purchase t can be approximated as a multiplicative process, resulting from independent random factors contributing to a reader's decision to buy a book. Such random multiplicative processes are shown to lead to a lognormal distribution [27–31].

Combining these mechanisms, we can write the probability $\Pi_i(t)$ of a book i to be purchased at a time t after publication as [26]

$$\Pi_i(t) \sim \eta_i S_i^t A_i(t), \tag{2}$$

which depends on (i) the book's fitness η_i, (ii) the total number of sales until t, S_i^t (preferential attachment), and (iii) the aging factor (1). Combining (i)–(iii), we find that the total sales of book i at time t after publication follows (a detailed derivation is given in Section S2.2 of the supplementary materials of [26]),

$$S_i^t = m\left[e^{\lambda_i \Phi(\frac{\ln t - \mu_i}{\sigma_i})} - 1\right], \tag{3}$$

where

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2}\, dy \tag{4}$$

is the cumulative normal distribution related to the error function as $\Phi(x) = 1/2\,\mathrm{erfc}(-x/\sqrt{2})$ where erfc is the complementary error function given by $1 - \mathrm{erf}(x)$ and λ_i is the relative fitness proportional to η_i.

To demonstrate how the model (2)–(4) can reflect actual sales, in Fig. 10(B) we show the sales pattern of *The Appeal* by John Grisham which sold over a quarter million copies in a single week after publication. We obtained the parameters $\lambda = 10.37$, $\mu = 2.03$ and $\sigma = 1.12$ by fitting Eq. (3) to the book's cumulative sales, the fit being shown in Fig. 10(C), trailing closely the real sales pattern ($R^2 = 0.99$). In fact, the model (3) can handily explain a wide range of sales patterns by varying only the three parameters μ, σ and λ (Fig. 10(E)).

A key prediction of model (3) is that by transforming all the sales curves into a single curve using rescaled variables, all sales curves should follow the same universal curve.

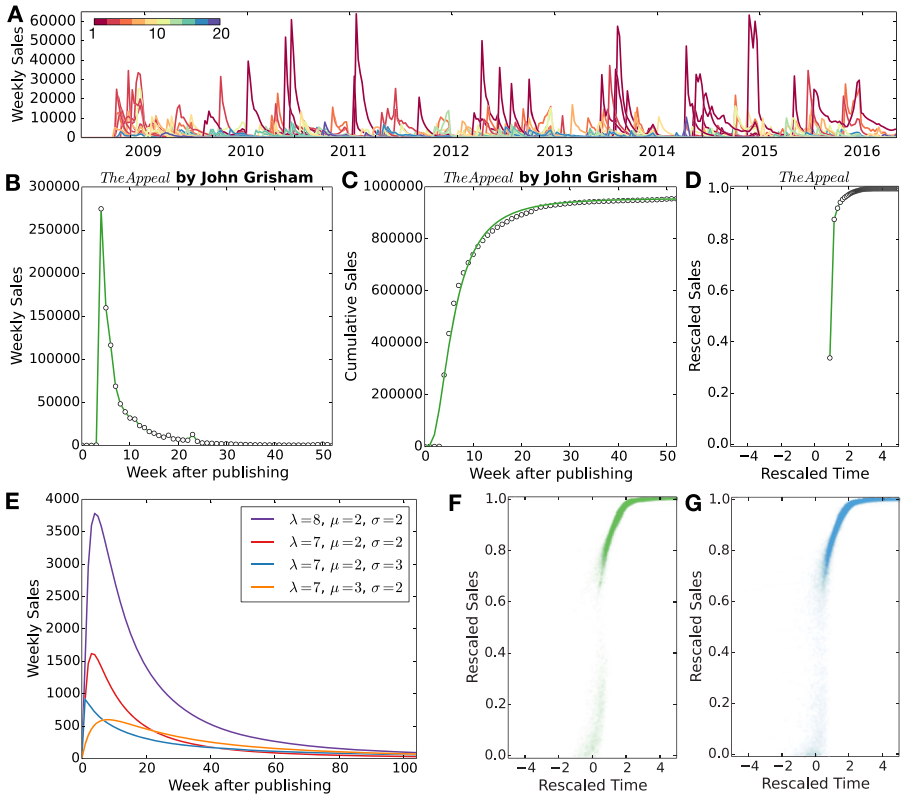

Figure 10 *Modeling sales patterns*. (**A**) Weekly sales of 200 randomly selected bestsellers, with colors indicating best rank on the NYTBL. (**B**) The weekly sale numbers of the bestselling hardcover *The Appeal* by John Grisham as reported by NPD Bookscan. (**C**) The fit on the cumulative sales of (**B**), empty circles showing the measured sales curve and the green line the model with parameters $\lambda = 10.37$, $\mu = 2.03$ and $\sigma = 1.12$. The fit accurately reflects the data, as evidenced by the high $R^2 = 0.99$ value. The only exception are the very first weeks where data is often unreliable due to discrepancies between the reported publication date and the book's physical availability in bookstores. (**D**) Rescaled sales curve for *The Appeal*. (**E**) How weekly sale curves change for different λ, σ and μ values. By tuning these, the many different types of sales histories shown in (**A**) can be accounted for. Rescaled sales curves of all hardcover bestsellers, (**F**) fiction and (**G**) nonfiction. Most book sales curves closely follow a single formula, indicating that the model captures the patterns correctly as given in Eq. (5)

These rescaled variables are $\tilde{t} \equiv (\ln t - \mu_i)/\sigma_i$ and $\tilde{S} \equiv \ln(1 + S_i^t/m)/\lambda_i$ and by substituting them into (3) we obtain

$$\tilde{S} = \Phi(\tilde{t}). \tag{5}$$

As an example, we show the rescaled curve for *The Appeal* in Fig. 10(D). The rescaled time $\tilde{t} = 1$ roughly corresponds to the time of the peak sale, and for this book there were almost no sales before that point.

If the model fits the sales pattern for all books, we expect the rescaled curves derived for all books to collapse into a single curve. We therefore measured the μ_i, σ_i and λ_i values for all books in the New York Times bestseller data using $m = 30$ and the Least-Square Fitting method on the available sales range for each book. We then rescaled the sales curve of each book accordingly, the rescaled sales curves being shown in Fig. 10(F) for fiction and (G) for nonfiction. The fact that all curves collapse into a single one indicates that the model correctly captures the sales pattern of most books.

One limitation of the proposed model is that it cannot account for exogenous events like awards, movie adaptations or mentions by prominent venues or celebrities. These events may land an otherwise unnoticed book on the New York Times bestseller list years after its original publication as we have seen in Figs. 9(C)–(F). Yet, these are exceedingly rare cases and most bestsellers follow a more typical sales pattern, one that is well accounted for by our model.

Taken together, we find that by using the fundamental mechanisms of fitness, preferential attachment and aging, we can explain and accurately model the sales curves of all bestsellers, regardless of genre. We can do this by relying on our observation that all books follow a well defined, regular path in selling copies including the timing of the peak sales, exceptions being rare.

5.3 Predicting future sales

In the previous section we have seen that only three parameters are needed to describe the sales history of any bestseller: the fitness λ, the immediacy μ and the decay rate σ. In Figs. 11(A)–(C) we show the probability distributions of each parameter for all bestsellers. In (A) we see that the fitness distribution $P(\lambda)$ is very similar for fiction and nonfiction bestsellers. This is to be expected, since these are all bestselling books and therefore all have high fitness. Yet, the variation of relative fitness is slightly higher for fiction

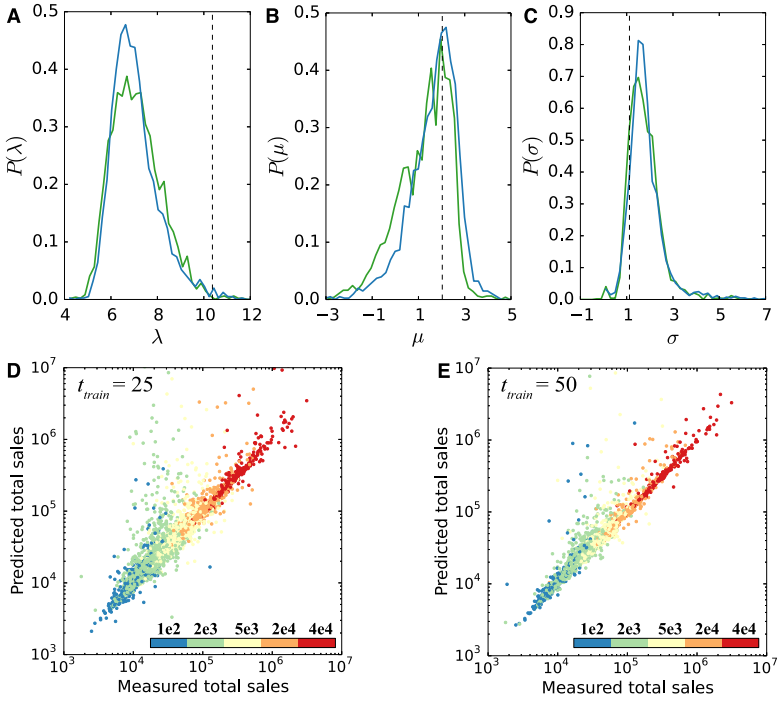

Figure 11 *Predicting sales*. Distribution of the parameters (**A**) fitness λ, (**B**) immediacy μ and (**C**) decay rate σ for all bestsellers, with fiction shown in green and nonfiction in blue. The dashed lines indicate the λ, μ and σ values of *The Appeal*, indicating high fitness resulting in a higher total sales, average immediacy indicating typical peak sales time and slightly lower than typical decay rate meaning a relatively slow drop in sales after the peak. Predicted vs measured total sales calculated from Eq. (6), using the first (**D**) 25 and (**E**) 50 weeks to obtain λ. The colors represent the peak sale number for each book. The first 6 months offers an accurate picture for the full total sales and the prediction accuracy increases as longer times are used to calculate the fit. Additionally, total sales of books with higher peak sales are predicted better than of those with lower peaks

than nonfiction, indicating a broader range. The observation that fiction bestsellers show more variability than nonfiction bestsellers is consistent with earlier findings about the one year (Fig. 3) and weekly (Fig. 4) sales. Additionally, the λ distribution for fiction peaks at a slightly higher value for fiction than nonfiction, indicating a higher relative fitness on average. This is because fiction books sell more copies than nonfiction books on average, as discussed in Sect. 3.

The relative fitness can singlehandedly predict how many copies a book will sell during its lifetime. Taking $t \to \infty$ in Eq. (3), we obtain [26]

$$S_i^{\infty} = m\left(e^{\lambda_i} - 1\right), \tag{6}$$

predicting that the total number of sales of a book in its lifetime depends only on a single parameter, the relative fitness λ. Consequently, if the model captures the data well, we expect a good match between the predicted and measured total sales, even when using data from a shorter time period than the book's lifetime to obtain the fitness parameter, allowing us to predict the total sales using (6). Results for different choices of time periods to calculate λ are shown in Fig. 11(D) for the first 25 weeks and (E) for the first 50 weeks after book release. We find that a fit derived from the first 25 weeks results in quite accurate predictions for the total sales of most books, indicating that our model can accurately predict how many copies a book will sell during its lifetime months after publication. As the number of weeks used for the fit increases, so does the accuracy of the prediction. Additionally, total sales of books with higher sales peaks are predicted more accurately, as indicated by the relative closeness of the red and orange dots to the 45 degree line as opposed to the green and blue dots which are generally more spread out.

Figure 11(B) shows the probability distribution for the immediacy parameter $P(\mu)$, indicating that both fiction and nonfiction books have similar immediacy distributions, i.e. all bestsellers reach their sales peak at a similar times. This result is consistent with Fig. 9, including the observation that $P(\mu)$ peaks at a slightly higher value for nonfiction than fiction, pointing to later peak sale times for nonfiction compared to fiction bestsellers.

Finally, the probability distribution for the decay rate values $P(\sigma)$ is shown in Fig. 11(C). The distributions for both fiction and nonfiction are quite narrow, following each other closely except at the very top, indicating very similar longevity and decay rates for all bestsellers. Yet, the distribution is slightly broader for fiction compared to nonfiction, indicating that on average, the longevity and continued success of fiction books vary more than nonfiction books, even among bestsellers.

The dashed lines on all three distributions show where the parameter values for *The Appeal* fall, indicating an extremely high fitness pointing to very high sales, typical immediacy pointing to average peak sale timing and lower than average decay rate pointing to a relatively slow drop in sales after the peak.

We find, overall, that the model (3) correctly describes the sales pattern of a book, accurately predicting the total sales once the book has been out for some time. However, we have seen in Fig. 9 that for the majority of bestsellers, most sales happen during the first few months. Consequently, a prediction of the future sales many months after the publication date is of value for inventory management.

6 Conclusions

The goal of this paper is to bring a big data perspective on the factors that influence book sales. For this, we developed a systematic, data driven approach to investigate the sales patterns of works and their creators that made it into the New York Times bestseller list.

We find that bestsellers have a higher chance of coming from the general fiction and memoir categories and regardless of the sub-genre, nonfiction books sell less copies than fiction books. In both categories, any book making it to the top of the bestseller list will sustain its sales longer compared to the books that barely make it to the list, indicating that the higher the initial success, the longer it will persist. There were no significant changes in the number of copies a book needs to sell in order to achieve bestseller status over the years since 2008, approximately the same amount of hardcovers being sold today as they were in the past years. This is a remarkable finding showing that the increasing availability of books in digital format has no influence on hardcover sales. Yet seasonal fluctuations within a year are important, influencing the relative success of a book compared to the rest of the market. Even though the holidays are times where substantially more books are sold, it is harder for any book to stand out because of these elevated sales.

From the author's perspective, we found fiction writers to be more prolific compared to nonfiction authors, achieving more repeat success. Such repeat success is helped by the serialized nature of many fiction bestsellers: When readers enjoy a series, subsequent books will have a higher potential of success. Interestingly, nonfiction authors writing in a serialized fashion focusing on a theme enjoy similar repeat success. As readers prefer the familiar over unknown, having some sense of what to expect drives more people towards a book or a series. This insight is consistent with the observation that people enjoy reading about celebrities or historic figures and events with whom they already have some degree of familiarity.

While gender disparity is prevalent in both academia and business, its largely absent in fiction: female and male authors are equally represented on the fiction bestseller list. In contrast, in nonfiction most bestsellers are written by male authors, showing that female authors either avoid nonfiction or are less successful if they do so. Yet, the breakdown of genres by gender and the finding that more romance is written by women and more mystery is written by men shows that stereotypical gender roles may be found in the world of authors as well.

Investigating the weekly sale numbers of bestsellers helped us identify a universal sales pattern: the sales increase very fast after a book's release, reaching their peak in the first ten weeks and drop dramatically afterwards. Using this universality we propose a statistical model that correctly describes the sales patterns of all bestsellers and accurately predicts the total number of copies an edition will sell during its lifetime a few months after a book's release, which could be of value for inventory management and assessing long-term impact. We find particularly interesting that a model originally proposed to describe citation patterns [26] offers an accurate description of book sales as well, albeit at different time scales. This suggests that the fundamental processes driving the attention economy of the phenomena—book selection and citations—are the same.

The discovery of the universal nature of sales patterns and its driving mechanisms are important for our understanding of the industry and how individuals buy books. In fact, the model (3) provides us with an excellent tool to reconstruct the sales timeline of any book from beginning to end, once the parameters λ, μ and σ are known. Combining that

with our insights about the characteristics of the bestsellers and their sales numbers, the ground is set for the development of tools to predict the parameters of the model before the book is published. Such a model could accurately foresee the entire sales curve of the given book months before that book is on the shelves, unlocking the full potential predictive power of the book industry.

We expect our findings on bestsellers to offer a starting point and inspiration to investigate the success of books and authors further, considering and comparing a variety of books including those that did not sell well, helping us to ultimately understand what it takes to be successful in an industry that is not only large and extremely competitive, but also effects us both as individuals and collectively as a society, by shaping our culture.

Acknowledgements
We wish to thank Douglas Abrams and Akos Erdös for giving us an insider's view to the publishing word and helping us understand it better, Kim Albrect and Alice Grishchenko for helpful visualizations, Peter Rupert and other colleagues at the CCNR, especially those in the success group, for valuable discussions and comments. A companion website with interactive visualisations can be found at http://bestsellers.barabasilab.com.

Funding
This research was supported by Air Force Office of Scientific Research (AFOSR) under agreement FA9550-15-1-0077, John Templeton Foundation under agreement 51977 and DARPA under agreement N66001-16-1-4067.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors designed the research. BY, XW and JH obtained, prepared and cleaned the data. BY and XW analyzed the data and prepared the figures. BY and A-LB prepared the manuscript. All authors read and approved the final manuscript.

Author details
[1]Center for Complex Network Research and Department of Physics, Northeastern University, Boston, USA. [2]CompleX Lab, Web Sciences Center, University of Electronic Science and Technology of China, Chengdu, China. [3]Center for Cancer Systems Biology, Dana Farber Cancer Institute, Boston, USA. [4]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA. [5]Center for Network Science, Central European University, Budapest, Hungary.

Publisher's Note

Received: 27 September 2017 Accepted: 12 January 2018 Published online: 06 April 2018

References

1. U.S. book industry/market—statistics & facts. Statista. https://www.statista.com/topics/1177/book-market/. Accessed 2015-09-29
2. Schmidt-Stölting C, Blömeke E, Clement M (2011) Success drivers of fiction books: an empirical analysis of hardcover and paperback editions in Germany. J Media Econ 24(1):24–47
3. Leemans H, Stokmans M (1992) A descriptive model of the decision making process of buyers of books. J Cult Econ 16(2):25–50
4. D'Astous A, Colbert F, Mbarek I (2006) Factors influencing readers' interest in new book releases: an experimental study. Poetics 34(2):134–147
5. Clement M, Proppe D, Rott A (2007) Do critics make bestsellers? Opinion leaders and the success of books. J Media Econ 20(2):77–105
6. Keuschnigg M (2015) Product success in cultural markets: the mediating role of familiarity, peers, and experts. Poetics 51:17–36
7. Nakamura L (2013) "Words with friends": socially networked reading on goodreads. Publ Mod Lang Assoc Am 128(1):238–243
8. Carmi E, Oestreicher-Singer G, Sundararajan A (2012) Is Oprah contagious? Identifying demand spillovers in online networks. NET Institute Working Paper 10-18
9. Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: online book reviews. J Mark Res 43(3):345–354
10. Tsur O, Rappoport A (2009) RevRank: a fully unsupervised algorithm for selecting the most helpful book reviews. In: ICWSM
11. Beck J (2007) The sales effect of word of mouth: a model for creative goods and estimates for novels. J Cult Econ 31(1):5–23
12. Ashok VG, Feng S, Choi Y (2013) Success with style: using writing style to predict the success of novels. Poetry 580(9):70
13. Johnson MW (2014) Bestsellers beyond bestsellers: the success of a good story. Online J Commun Media Technol 4(4):1

14. Sorensen AT, Rasmussen SJ (2004) Is any publicity good publicity? A note on the impact of book reviews. NBER Working Paper, Stanford University
15. Clerides SK (2002) Book value: intertemporal pricing and quality discrimination in the US market for books. Int J Ind Organ 20(10):1385–1408
16. Kovács B, Sharkey AJ (2014) The paradox of publicity: how awards can negatively affect the evaluation of quality. Adm Sci Q 59(1):1–33
17. Sorensen AT (2007) Bestseller lists and product variety. J Ind Econ 55(4):715–738
18. Verboord M (2011) Cultural products go online: comparing the Internet and print media on distributions of gender, genre and commercial success. Communications 36(4):441–462
19. Verboord M (2012) Female bestsellers: a cross-national study of gender inequality and the popular–highbrow culture divide in fiction book production, 1960–2009. Eur J Commun 27(4):395–409
20. Deschâtres F, Sornette D (2005) Dynamics of book sales: endogenous versus exogenous shocks in complex networks. Phys Rev E 72(1):16112
21. About the best sellers. New York Times. http://www.nytimes.com/books/best-sellers/methodology/. Accessed 2014-09-29
22. Krystal A (2012) EASY WRITERS: guilty pleasures without guilt. The New Yorker
23. Grossman L (2012) Literary revolution in the supermarket aisle: Genre fiction is disruptive technology. Time
24. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
25. Caldarelli G (2007) Scale-free networks: complex webs in nature and technology. Oxford University Press, London
26. Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342(6154):127–132
27. Boag JW (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J R Stat Soc, Ser B, Methodol 11(1):15–53
28. Sartwell PE et al (1950) The distribution of incubation periods of infectious disease. Am J Hyg 51:310–318
29. Preston FW (1981) Pseudo-lognormal distributions. Ecology 62(2):355–364
30. Williams CB (1940) A note on the statistical analysis of sentence-length as a criterion of literary style. Biometrika 31(3/4):356–361
31. Herdan G (1958) The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. Biometrika 45(1–2):222–228