A time series analysis of Amazon sales rank

Eric Blair

10 August 06

I have been very interested in the sales of *Math You Can't Use: Patents, Copyright, and Software*, a book with which I was heavily involved. (Amazon page)

So naturally, I've been tracking the Amazon sales rank. At first, I did it the way everybody else does—refreshing the darn page every twenty minutes—but I have recently started doing it the civilized way—an automated script. Here is what I've learned about how Amazon does its rankings.

Background and conclusion

First, to give you some intuition as to sales rank, here's a little table:

1-10	Oprah's latest picks
10-100	The NYT's picks
100-1,000	Books by editors of Wired Magazine, topical rants by pundits/journalists, 'classics'
1,000-500,000	everything else (still selling)
500,000-2mil	everything else (technically in stock)

How much more detail can we get? The answer: none, really. You'll see below that over the course of a few days, the ranking of a typical book will go from 50,000 to 500,000, and a minute later it will be back at 50,000. Thus, the sort of things we usually do with a ranking, like compare two books, are unstable to the point of uselessness.

One thing you evidently can do with the ranking is determine whether a book has sold a copy in the last hour or two. As you'll see below, there's a simple formula that will work for most books: if (current rank) > (earlier rank) then there was a recent sale.

The first chart

Here is a graph of sales of *Math You Can't Use*. A data point is added every three hours, so if you come back in a week, this graph will be different. See below for the code I used to generate this. On the x-axis is the date, and on the y-axis is the Amazon sales rank at that time.

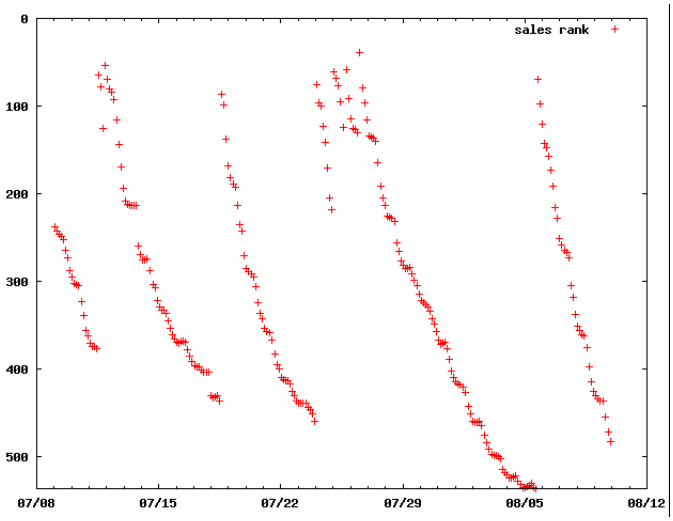

Figure 1: Sales rank for Klemens's *Math You Can't Use*

You can see from the graph that the pattern is a sudden jump and then a slow drift downward. The clearest explanation is that the sales rank is basically a function of last sale. When a copy sells, the book jumps to a high rank, and then gets knocked down one unit every time any lower-ranked book sells.

There are lots of details that those of us not working at Amazon will never quite catch. There are periods (sometimes mid-day) when the rank drifts down more slowly than it should, then speeds up in its descent. This implies to me some computational approximations that eventually get corrected. You'll notice that some of the books below show a small slope upward (a ten or twenty point rise in ranking) from time to time. When this happens, lots of books do it at once, also indicating some sort of correction whose purpose or method I don't have enough information to divine. Epstein and Axtell's book rises appreciably when it nears half a million. Filally, I don't have enough data to determine whether the ranking distinguishes between sales of used and new copies; I don't think it does.

Others

In the online version of this analysis, you'll find the sales ranks over time for some other books. They're dynamically updated graphs, so they're better off online than on paper.

Online, the reader will also find the code that generated these graphs, for tracking the reader's own favorite books. Again, code for cutting and pasting works best on screen.

Executive summary

At this point, I'm not sure why Amazon ranks books below the top thousand, except for a sort of geek factor. For all of the books here, it is basically impossible to say something like '*Tivo for Dummies* is ranked around 100,000,' since the ranking jumps by an order of magnitude almost daily. Similarly, there's no point saying '*Atonement* is ranked higher than *Great Expectations*', since you have a 50-50 chance of being wrong tomorrow. All we get is a very broad ballpark figure (a football field figure?), and a too-good impression of how many hours ago the last sale was made.

Those of us interested in the sales rank of books outside Oprah's picks would be better served if the system were less volatile. In technical terms, if my guess that the score experiences exponential decay is correct, then the ranking system would be more useful to those of us watching the long tail if the decay factor were set to a smaller value.

Technical notes

The data looks to me like an exponential decay system, where you have a current score S_t which goes up by some amount every sale, but drifts down by some discount rate every period, $S_{t+1} = \lambda S_t$. [Thus, if there were no sales events, your score would be $S_t = S_0 \exp(-\lambda t)$.]

To fit this, I flipped and renormalized the rankings so that one was the highest possible ranking, and zero corresponded to a ranking of 500,000. Then, I set the following algorithm:

- The score was initialized at 0.58.
- Each period, score is multiplied (shrinks) by a factor of 0.96.
- If there is a sale, then score rises by the addition of (1-current score) * 0.79.

As you can imagine, I found those constants via minimizing the distance between the estimate and the actual. The algorithm is an exponential decay model with $\lambda = 0.96$, and upward shocks as described. The only way I could fit the data was to make shocks when the book is at a low sales rank bigger than shocks when it has a high sales rank. There's surely a more clever way to do it.

The green line shows the exponential decay model fit to the actual data. You can decide if this is a good fit or a lousy one.

You can also have a look at how the model fit to Madonna's book.

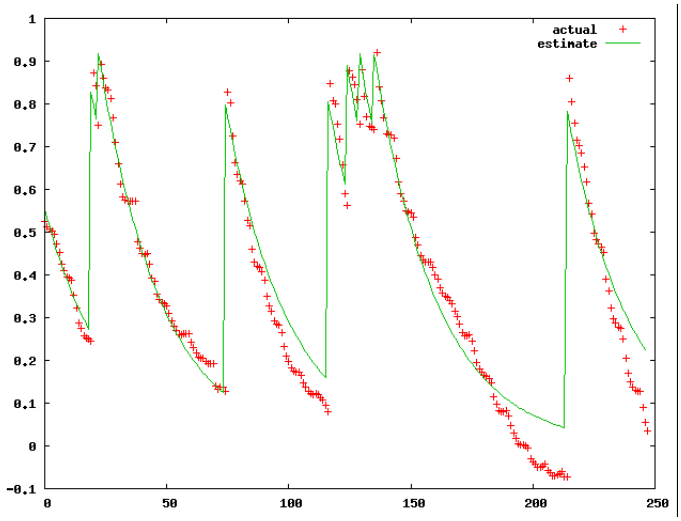

Figure 2: My attempts to fit the Amazon sales rank to an exponential model