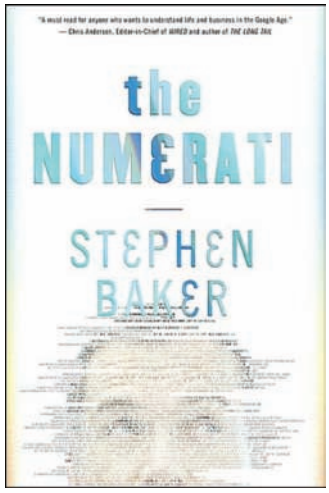
The Numerati

Reviewed by Jeffrey Shallit



The Numerati
Stephen Baker
Houghton Mifflin Co., 2008
US$26.00, 256 pages
ISBN-13: 978-0618784608

Is it possible for a nonmathematician to write both accurately and entertainingly about a mathematical topic while still conveying something nontrivial about the mathematics? The answer is yes, but good examples are rare. Constance Reid did the trick with her book about Hilbert, and, to a lesser extent, with her book about Courant, but she had the advantage of having Julia Robinson for a sister. And of course Martin Gardner, who had little formal mathematical training, wrote the "Mathematical Games" column of *Scientific American* for many years, and introduced the beauty of mathematics to many young readers, including this reviewer.

Stephen Baker, the author of *The Numerati*, is, unfortunately, no Martin Gardner. By *numerati* (the word apparently first appeared in a 1990 review of a British art exhibit, written by Doron Swade) Baker means the kind of people who, were they working in the financial industry, would be called "quants": people with very strong mathematical and computer skills who can analyze real-world problems. While "quants" study financial markets and build mathematical models, Baker's *numerati* analyze large volumes of data collected electronically, in order to make predictions about human behavior in a variety of spheres: voting, employment, consumption, crime, illness, blogging, and marriage. Each of these activities gets a chapter devoted to it, in which Baker interviews several people who analyze the relevant data and try to come up with marketable conclusions. "[T]hese mathematicians and computer scientists," Baker intones sternly, "are in a position to rule the information of our lives."

In a book whose subject is data, equations, and mathematical models, Baker is surprisingly shy about presenting any actual mathematics. Or perhaps it is not so surprising. Steven Hawking once wrote, "Someone told me that each equation I included in the book would halve the sales. I therefore resolved not to have any equations at all. In the end, however, I did put in one equation, Einstein's famous equation $E = mc^2$. I hope that this will not scare off half of my potential readers." [2] Baker has taken Hawking one equation further.

Baker's approach is almost entirely anecdotal. All told, he interviews about two dozen of the *numerati*, ranging from IBM's Samer Takriti to Yahoo's head of research, Prabhakar Raghavan, and asks them some not-very-revealing questions. We learn very little about their personalities and even less about what it is they do on a day-to-day basis.

Some of the anecdotes are, admittedly, interesting. I particularly enjoyed the plan to "put a wireless computer on half a million cows in Kansas"; with the data collected, researchers hope to determine what behavior patterns of cows are correlated with higher-quality meat. But some are not so interesting. Baker opens with a puzzle: why do people who rent romantic movies online also tend to click on an ad for rental cars, much more than the average user? The answer, when it comes, is not that surprising: lovers of romantic movies were attracted by the ads that promoted weekend "escapes".

There is very little in *The Numerati* to interest the professional or amateur mathematician; this is the kind of book that a business executive

Jeffrey Shallit is professor of computer science at the University of Waterloo, Ontario, Canada. His email address is shallit@cs.uwaterloo.ca.

might buy in an airport bookstore, hoping to learn something about mathematical modeling and the Internet—but I imagine even the business executive will find insufficient novelty in Baker's modest survey. There's just not enough detail provided to tell the reader very much about the main subject: the models and algorithms that extract meaning from large volumes of data.

As an example, consider this passage: "If one of Raghavan's scientists gives an imprecise computer command while trawling through Yahoo's data, he can send the company's servers whirring madly through the noise for days on end. But a timely tweak in these instructions can speed up the hunt by a factor of 30,000. That reduces a 24-hour process to about three seconds. His point is that people with the right smarts can summon meaning from the nearly bottomless sea of data. It's not easy, but they can find us there."

Reading this, I can only wonder, what is an "imprecise computer command"? Does the passage concern a new breakthrough at Yahoo in search optimization, or something obvious that every undergraduate computer science student learns, such as binary search? Baker just doesn't give enough detail to decide.

Baker emphasizes that the volume of data collected by the *numerati* requires new techniques, but he doesn't really explain why. It would have been nice to read something along these lines: if we are working with small data sets, with hundreds or thousands of items, we can afford to use algorithms that run in linear, $O(n \log n)$, or even quadratic time. But, as my colleague Alex López-Ortiz has noted [3], when you are dealing with 2^{30} or even 2^{40} data points, the log factor is the difference between a query that completes in a second and one that completes in half a minute.

Too often Baker relies on clichés. Over and over, we are told that the goal of the *numerati* is to "turn us into dizzying combinations of numbers" (p. 13), to "turn IBM's workers into numbers" (p. 20), and that they will view people as "boiled down to numbers" (p. 23) or "represented as a series of numbers" (p. 35). Of these, only the last is accurate. Sometimes, though, Baker says we are actually equations: "each of us [is] represented by scores of equations" (p. 42); "I had ... no clue as to what kind of equation I would become" (p. 99). This, even metaphorically, seems incorrect. People might be represented by numbers, and their relationships might be governed by equations, but it makes little sense to claim that an individual's attributes are represented by an equation.

Although most of his account is accurate—as far as it goes—Baker does get some of the history wrong. He claims, for example, that "Google's breakthrough, which transformed a simple search engine into a media giant, was the discovery that our queries—the words we type when we hunt for Web pages—are of immense value to advertisers". This is incorrect. The site `Goto.com` allowed advertisers to bid on search results as early as February 1998, two years before Google did so. Google's original noteworthy accomplishment—and the one that made it the search engine of choice—was its new algorithm, called PageRank, for deciding what Web pages provide good matches for a query.

PageRank represented the Web as a directed graph. Nodes are pages, and there's a directed edge from page A to page B if A links to B. In its simplest form, PageRank assigned a weight W to the edge (A, B) with

$$W = \frac{\text{number of links from } B \text{ to } A}{\text{total number of pages that } B \text{ links to}}.$$

The resulting square matrix, called the "link matrix", is column stochastic and has an eigenvalue of 1. The associated eigenvector, if it is unique and suitably normalized, gives the "rank" or importance of each page. (There is now more actual mathematics in this review than in all 244 pages of Baker's book.) To make this idea work well in practice, we need uniqueness of the eigenvector and a fast way to calculate it, so the mathematical story doesn't end here. But even in its infancy, PageRank helped Google give much better results than other search engines—so good that Google's home page cockily offers an option labeled "I'm Feeling Lucky", where only a single search result, the top one, is revealed—that it quickly became the search engine of choice. Although Google's search engine has since moved far past PageRank, a mathematically savvy writer could have easily summarized these elementary ideas, or at least referred to the paper of Bryan and Leise [1].

Even when a simple geometric diagram would have enlightened the reader, Baker refuses to provide it. In talking with Mark Steitz, a Democratic consultant, he describes a "simplex triangle" that represents voters in an election. Each voter is represented by a point with two coordinates that represent (a) the likelihood of favoring one party over another and (b) the likelihood of actually going to the polls in any election. "Steitz draws a vertical line up the triangle, a so-called isoquant. Each voter along this line is of equal value, he says." Although I imagine every reader of this review could produce the diagram Steitz has in mind, one picture here would be worth more than a hundred words.

In the chapter on politics, Baker discusses the difficulty of obtaining good data on who people are likely to vote for. Because of this, "proxies" are used; if you bought a Volvo and shop at Trader Joe's, you might be more likely to vote for a Democrat than someone who's an NRA member and drives a pickup truck. Geographical proxies can be good predictors, too, but Baker's account

is superficial compared to others, such as Michael Weiss's *The Clustering of America* [5].

The contrast between this book and some related ones published recently is startling. For example, Emanuel Derman's *My Life as a Quant* [4] is a memoir of the author's career as a physicist, computer programmer, and financial wizard. Along the way, Derman provides portraits of Tsung-Dao Lee, the physicist who co-discovered the asymmetry of the weak interaction with C. N. Yang, and Fischer Black, co-creator of the Black-Scholes equation for the value of an option. Here is Derman on T. D. Lee:

> ... every speaker felt compelled to focus on him; as they spoke, their eyes fixated only on him, and he let no statement he did not fully agree with pass him by. No matter who lectured at the seminar, T. D. concentrated intensely on their argument, and interrupted at the first instant something was not satisfactory. At times he broke in on the initial sentence of the talk, refusing to let a speaker proceed until the point was clarified. Sometimes clarification never came; I once witnessed the humiliation of a visting postdoc who was forced to defend the first sentence he uttered for the entire hour and a half allowed for his seminar.

Derman's writing is witty, insightful, and moving; his prose is eloquent, and accurately captures the joys and sorrows of doing research. Derman's book is not filled with equations, either, but he uses diagrams effectively to make his points, and describes, in a clear if nontechnical way, some of the ideas that excited him in physics and finance. As someone who has actually worked in mathematics, physics, and finance, Derman writes with an authority and insight that Baker cannot approach.

Very little of *The Numerati* is devoted to an analysis of the ethical and privacy concerns that data collection raises. Although Baker briefly discusses one way of hiding from the *numerati*—an initiative called Attention Trust—he says almost nothing about technologies for cryptography and anonymity. Modern cryptography, which is strongly mathematically based, offers us the hope that many of our transactions can take place veiled from the prying eyes of the *numerati*. And anonymous Web-surfing, based (for example) on technology from `anonymizer.com` or the Tor project, can prevent data collectors from linking online behavior with the specific person who is doing the surfing.

Ultimately, I did not find *The Numerati* a very satisfying account of its subject. I wanted more insight—something that Baker, with his nonmathematical background, could not provide. Perhaps I am unfair in criticizing Stephen Baker for not writing the book I would have wanted to read. The problem is, I don't think he wrote the book that most people would have wanted to read.

References

[1] KURT BRYAN AND TANYA LEISE, The $25,000,000,000 eigenvector: The linear algebra behind Google, *SIAM Review* **48** (2006), 569–581.

[2] STEVEN HAWKING, *A Brief History of Time*, Bantam Books, 1998.

[3] ALEJANDRO LÓPEZ-ORTIZ, Algorithmic foundations of the Internet, *Combinatorial and Algorithmic Aspects of Networking*, Lecture Notes in Computer Science, Vol. 3405, Springer, Berlin, 2005, pp. 155–158.

[4] EMANUEL DERMAN, *My Life as a Quant: Reflections on Physics and Finance*, Wiley, 2004.

[5] MICHAEL J. WEISS, *The Clustering of America*, Tilden Press, 1988.