

The Matthew Effect & Search Results

<http://crookedtimber.org/2009/09/19/the-matthew-effect-search-results/>

by John Holbo on September 19, 2009

Some thoughts, related to [Michael's 'going pro' post](#) and [Kieran's recent post](#) on impact factor. To what extent is the whole internet afflicted with [the Matthew Effect](#)? “For to all those who have, more will be given, and they will have an abundance; but from those who have nothing, even what they have will be taken away.” If you want to be a bit more specific, to what degree are search results afflicted by it?

Let me illustrate with a couple cases I've personally noted, which I suspect are representative. I just wrote [a book about Plato](#) [update: now optimized!], so naturally I'm curious what comes up if you Google [Plato](#). Predictably: Wikipedia. The Stanford Encyclopedia of Philosophy. (I'm going to ignore erroneous results, due to ambiguity: computer systems named Plato, famous drivers named Plato, former child star actresses who committed suicide named Plato.) You get somewhat arbitrary Google book results. Why, in particular, is an edition of the *Theaetetus*, edited by Robin Waterfield #6? You also get a number of pages that, not to put too fine a point on it, look to have been designed along 1996-1999 lines. Because that's surely when they were originally posted. [This page](#), for example, is #2, right after Wikipedia, beating out even the SEP. Now, that's nuts. There's nothing wrong with the page, as far as it goes. But it's clearly a beneficiary of the Matthew Effect. Google users are brought to this page – in droves, I'll wager – because it was posted by an early-adopter of the interwebs thingummy. A similar example is [this page](#), coming in at #6. This one is a much more serious project, by someone who is clearly competent to write about Plato, and who moreover has worked pretty hard to maintain and build-up this site. (Not that I'm implying the author of that other page was not competent. Just that the content hardly explains the #2 ranking.) That second site posts public stats, which are interesting: “870 000 visits in 2008 (an average of 2 374 visits per day).” I wouldn't be surprised if the author of this site is, in a way, the world's most influential Plato scholar, due to the fact that he had the good luck to start posting in 1996. Out of the top 10 hits for Plato (ignoring erroneous hits) we get, by my count: 2 that clearly deserve to be in the top 10 – Wikipedia and the SEP; 3 Google Books titles that are perfectly respectable but pretty random – i.e. none of the three is one of the first titles you would mention to someone asking ‘where should I start, to find out about Plato?’; 3 personally-maintained sites that are clearly here because they are late-1990's Matthew Effect beneficiaries; a [pretty good animated video](#) of the Cave Parable on YouTube; and a link to [the Plato page](#) of the MIT Classics Archive, which – despite the academic imprimatur – is a late 1990's affair. Another Matthew case. (The last time I visited, a lot of the links were broken. But maybe someone has fixed that.) The content is Jowett translations; that is, old stuff.

What are we missing? [The Perseus Project](#), for one. I was surprised to see no Amazon links cracking the top 10. (Not that I think that's so important, but I'm surprised.) How did we do? So-so. Partly the problem is that you should enter more intelligent search parameters. But part of the problem is runaway Matthew Effect. I suspect that the three random book hits could be explained by the Matthew Effect, in some way. Someone must have linked to these books. And these titles,

rather than some others, lucked into a high slot. It's interesting that Google doesn't do better. (Not that I have any bright ideas.)

Second case: last year I posted [this X-Mas card set on Flickr](#). (I'm making more this year!) Anyway, long story short, one of the images got [Stumbled](#), as a result of which, eventually, two rather similar images diverged dramatically in their traffic. [This one](#) has been viewed 2,000 times. [This one](#) has been seen 12,000 times and has thereby accounted for 10% of the traffic my Flickr account has ever received. (I actually like the first image better.) I was curious whether it would just go and go like that forever, but recently it's stopped. My Stumblejuice ran dry. (The part of me that values justice is glad to see this. The part of me that likes getting free stuff for no good reason is a bit dismayed.) Anyway, I don't really understand how ranking sites like Stumble and Delicious and Digg and so forth work because I don't use them myself. But it strikes me that all this stuff clutters things up worse, Matthew-wise. [UPDATE: clarification. I don't mean the one pic got a huge spike that then disappeared. I've gotten those, too. Rather, you get a steady, slightly higher rate of traffic – in my case, 25-50 hits a day for months and months and months. But all that adds up.]

What could search engines do to combat the Matthew Effect better, algorithmically? Obviously if anyone knew, then Google would know, and presumably Google would then do it. (Or would they?)

[Jakob Nielsen's Alertbox, October 9, 2006:](#)

Participation Inequality: Encouraging More Users to Contribute

http://www.useit.com/alertbox/participation_inequality.html

Summary:

In most online communities, 90% of users are lurkers who never contribute, 9% of users contribute a little, and 1% of users account for almost all the action.

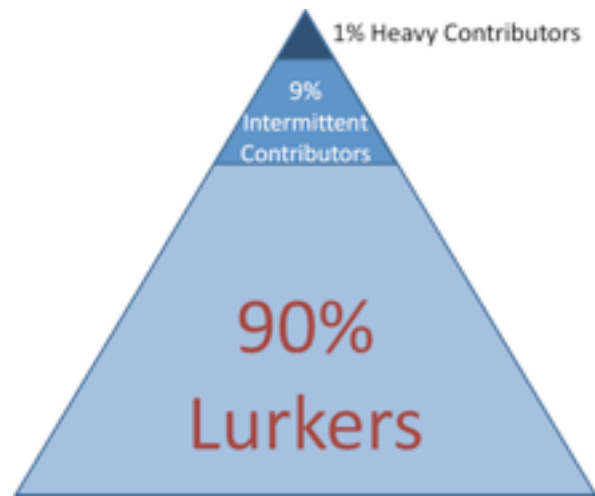
All large-scale, multi-user communities and online social networks that rely on users to contribute content or build services share one property: **most users don't participate** very much. Often, they simply **lurk** in the background.

In contrast, a tiny minority of users usually accounts for a disproportionately large amount of the content and other system activity. This phenomenon of **participation inequality** was first studied in depth by Will Hill in the early '90s, when he worked down the hall from me at Bell Communications Research (see references below).

When you plot the amount of activity for each user, the result is a [Zipf curve](#), which shows as a straight line in a [log-log diagram](#).

User participation often more or less follows a **90-9-1 rule**:

- **90%** of users are **lurkers** (i.e., read or observe, but don't contribute).
- **9%** of users contribute **from time to time**, but other priorities dominate their time.
- **1%** of users participate a lot and **account for most contributions**: it can seem as if they don't have lives because they often post just minutes after whatever event they're commenting on occurs.



Early Inequality Research

Before the Web, researchers documented participation inequality in media such as Usenet newsgroups, CompuServe bulletin boards, Internet mailing lists, and internal discussion boards in big companies. A study of more than 2 million messages on Usenet found that 27% of the postings were from people who posted only a single message. Conversely, the most active 3% of posters contributed 25% of the messages.

In Whittaker et al.'s Usenet study, a randomly selected posting was equally likely to come from one of the 580,000 low-frequency contributors or one of the 19,000 high-frequency contributors. Obviously, if you want to assess the "feelings of the community" it's highly unfair if one subgroup's 19,000 members have the same representation as another subgroup's 580,000 members. More importantly, such inequities would give you a **biased understanding of the community**, because many differences almost certainly exist between people who post a lot and those who post a little. And you would never hear from the silent majority of lurkers.

Inequality on the Web

There are about [1.1 billion Internet users](#), yet **only 55 million users (5%) have weblogs** according to Technorati. Worse, there are only [1.6 million postings per day](#); because some people post multiple times per day, **only 0.1% of users post daily**.

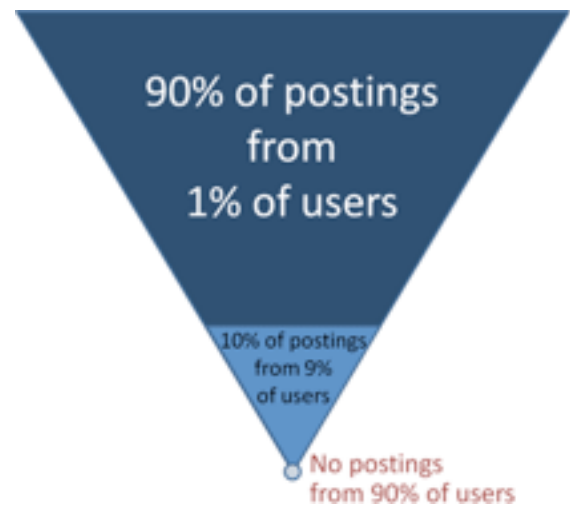
Blogs have even worse participation inequality than is evident in the 90-9-1 rule that characterizes most online communities. With blogs, the rule is more like 95-5-0.1.

Inequalities are also found on Wikipedia, where more than 99% of users are lurkers. According to [Wikipedia's "about" page](#), it has only 68,000 active contributors, which is **0.2%** of the 32 million unique visitors it has in the U.S. alone.

Wikipedia's most active 1,000 people — 0.003% of its users — contribute about two-thirds of the site's edits. Wikipedia is thus even more skewed than blogs, with a 99.8-0.2-0.003 rule.

Participation inequality exists in many places on the Web. A quick glance at Amazon.com, for example, showed that the site had sold thousands of copies of a book that had only 12 reviews, meaning that **less than 1% of customers contribute reviews**.

Furthermore, at the time I wrote this, 167,113 of Amazon's book reviews were contributed by just a few "[top-100 reviewers](#)"; the most prolific reviewer had written 12,423 reviews. How anybody can write that many reviews — let alone read that many books — is beyond me, but it's a classic example of participation inequality.



Downsides of Participation Inequality

Participation inequality is not necessarily unfair because "some users are more equal than others" to misquote *Animal Farm*. If lurkers want to contribute, they are usually allowed to do so. The problem is that the overall system is **not representative** of average Web users. On any given user-participation site, you almost always hear from the same 1% of users, who almost certainly differ from the 90% you never hear from. This can cause trouble for several reasons:

- **Customer feedback.** If your company looks to Web postings for customer feedback on its products and services, you're getting an unrepresentative sample.
- **Reviews.** Similarly, if you're a consumer trying to find out which restaurant to patronize or what books to buy, online reviews represent only a tiny minority of the people who have experiences with those products and services.
- **Politics.** If a party nominates a candidate supported by the "netroots," it will almost certainly lose because such candidates' positions will be too extreme to appeal to mainstream voters. Postings on political blogs come from less than 0.1% of voters, most of whom are hardcore leftists (for Democrats) or rightists (for Republicans).
- **Search.** Search engine results pages (SERP) are mainly sorted based on how many other sites link to each destination. When 0.1% of users do most of the linking, we risk having search relevance get ever more out of whack with what's useful for the remaining 99.9% of users. Search engines need to rely more on behavioral data gathered across samples that better represent users, which is why they are building Internet access services.
- **Signal-to-noise ratio.** [Discussion groups drown in flames](#) and low-quality postings, making it hard to identify the gems. Many users stop reading comments because they don't have time to wade through the swamp of postings from people with little to say.

Skewed Lurker–Contributor Ratio for Non-Profit Social Network

(Update 2009) The "Causes" application on Facebook had **25 million users** in April 2009, but only **185,000 had given a donation**, even though the application offers the ability to give to 179,000 different non-profit organizations. (This according to the [Washington Post](#).) Thus, social networking for charity fundraising has a **99.3% lurkers and 0.7% contributors** rule — even more skewed than the other participation inequalities we have seen. The data doesn't say how many of the 0.7% of users who donated have been *frequent* contributors, but most likely it's less than 1/10, meaning that the full rule would look something like **99-1-0**.

This finding comes as no big surprise, for three reasons:

- Despite the hype, Facebook is just another form of collaborative environment, meaning that long-established laws for online communities should hold. Maybe with small modifications, but the basics are due to human nature and don't change when moving to a new platform.
- Donating money is a stronger form of action than simply writing user-contributed content, so it makes sense that this form of contribution would have extremely strong participation inequality. If we measured the amount of money donated and not just a binary give/not-give distinction, the skew would likely be even more extreme.
- Our research on the [user experience of donating to charities online](#) found that most non-profits don't provide the information users want before they're willing to be separated from their money. (Or the info isn't shown in a sufficiently Web-oriented manner.)

How to Overcome Participation Inequality

You can't.

The first step to dealing with participation inequality is to recognize that it will always be with us. It's existed in every online community and multi-user service that has ever been studied.

Your only real choice here is in how you shape the inequality curve's angle. Are you going to have the "usual" 90-9-1 distribution, or the more radical 99-1-0.1 distribution common in some social websites? Can you achieve a more equitable distribution of, say, 80-16-4? (That is, only 80% lurkers, with 16% contributing some and 4% contributing the most.)

Although participation will always be somewhat unequal, there are ways to better equalize it, including:

- **Make it easier to contribute.** The lower the overhead, the more people will jump through the hoop. For example, Netflix lets users rate movies by clicking a star rating, which is much easier than writing a natural-language review.
- **Make participation a side effect.** Even better, let users participate with zero effort by making their contributions a side effect of something else they're doing. For example, Amazon's "*people who bought this book, bought these other books*" recommendations are a side effect of people buying books. You don't have to do anything special to have your book preferences entered into the system. Will Hill coined the term **read wear** for this type of effect: the simple activity of reading (or using) something will "wear" it down and thus leave its marks — just like a cookbook will automatically fall open to the recipe you prepare the most.

- **Edit, don't create.** Let users build their contributions by modifying existing templates rather than creating complete entities from scratch. Editing a template is more enticing and has a gentler learning curve than facing the horror of a blank page. In avatar-based systems like Second Life, for example, most users modify standard-issue avatars rather than create their own.
- **Reward — but don't over-reward — participants.** Rewarding people for contributing will help motivate users who have lives outside the Internet, and thus will broaden your participant base. Although money is always good, you can also give contributors preferential treatment (such as discounts or advance notice of new stuff), or even just put gold stars on their profiles. But don't give too much to the most active participants, or you'll simply encourage them to dominate the system even more.
- **Promote quality contributors.** If you display all contributions equally, then people who post only when they have something important to say will be drowned out by the torrent of material from the hyperactive 1%. Instead, give extra prominence to good contributions and to contributions from people who've proven their value, as indicated by their [reputation ranking](#).

Your website's design undoubtedly influences participation inequality for better or worse. Being aware of the problem is the first step to alleviating it, and finding ways to broaden participation will become even more important as the Web's social networking services continue to grow.

References

Laurence Brothers, Jim Hollan, Jakob Nielsen, Scott Stornetta, Steve Abney, George Furnas, and Michael Littman (1992): "Supporting informal communication via ephemeral interest groups," *Proceedings of CSCW 92, the ACM Conference on Computer-Supported Cooperative Work* (Toronto, Ontario, November 1-4, 1992), pp. 84-90.

William C. Hill, James D. Hollan, Dave Wroblewski, and Tim McCandless (1992): "Edit wear and read wear," *Proceedings of CHI'92, the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, CA, May 3-7, 1992), pp. 3-9.

Steve Whittaker, Loren Terveen, Will Hill, and Lynn Cherny (1998): "The dynamics of mass interaction," *Proceedings of CSCW 98, the ACM Conference on Computer-Supported Cooperative Work* (Seattle, WA, November 14-18, 1998), pp. 257-264.