

A response to "Scaling Clay Shirky"

(<http://www.theisociety.net/archives/000433.html>)

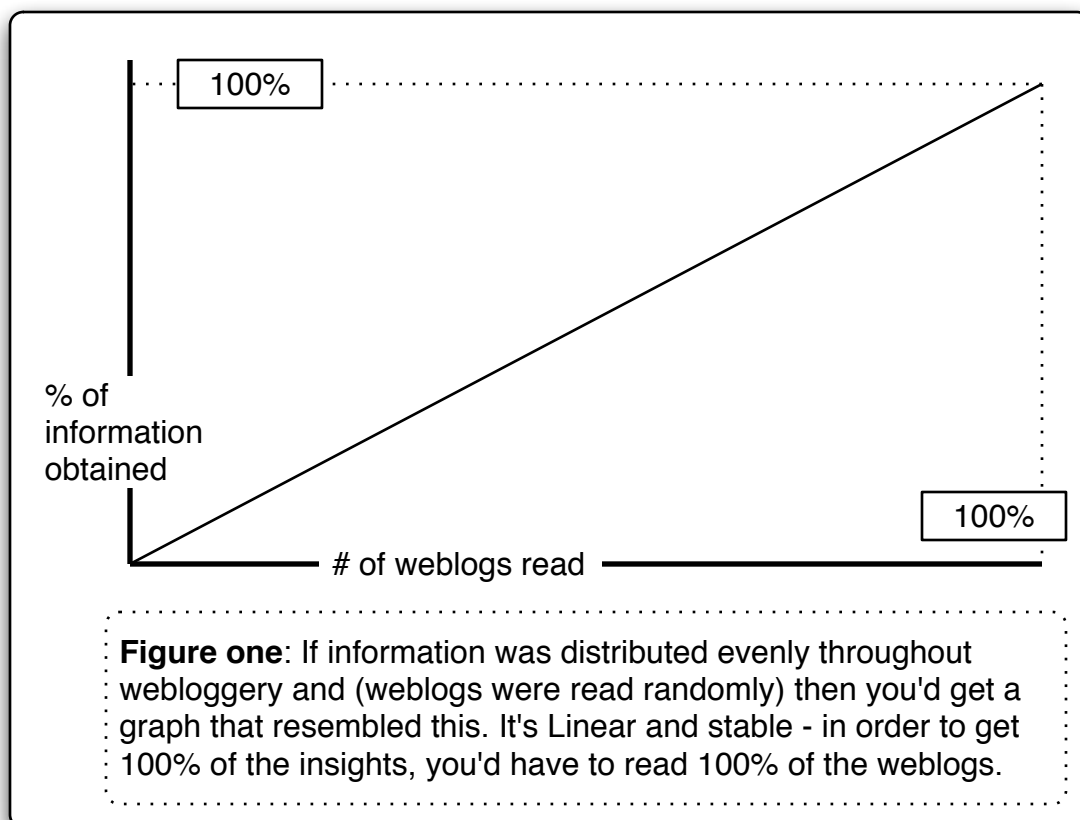
This piece originally appeared on [plasticbag.org](http://www.plasticbag.org):

http://www.plasticbag.org/archives/2003/05/how_do_we_find_information_in_the_blogosphere.shtml

"Clay's only crime is that he wrote an interesting article, about which there are many, many things to say. Consequently, lots of people wanted to say things about it. And because they all ran blogs, they did.

A predictable pattern soon emerged. In no time at all there were far too many commentary posts for anyone to read them all. Compounding this is the fact that with so many posts appearing on small, poorly linked sites, many comments were repeated. And each person who posted in ignorance something already said elsewhere muddied the waters further."

Let's aggregate all the insights about a given subject from all the weblogs that refer to it. This total aggregation will represent 100% of the information available on this subject.



However, we know it to be the case that information will not be distributed evenly throughout these weblogs. Many weblogs will contain limited information of any kind. Some will contain a lot. Many will contain replicated information that could easily be found on other sites.

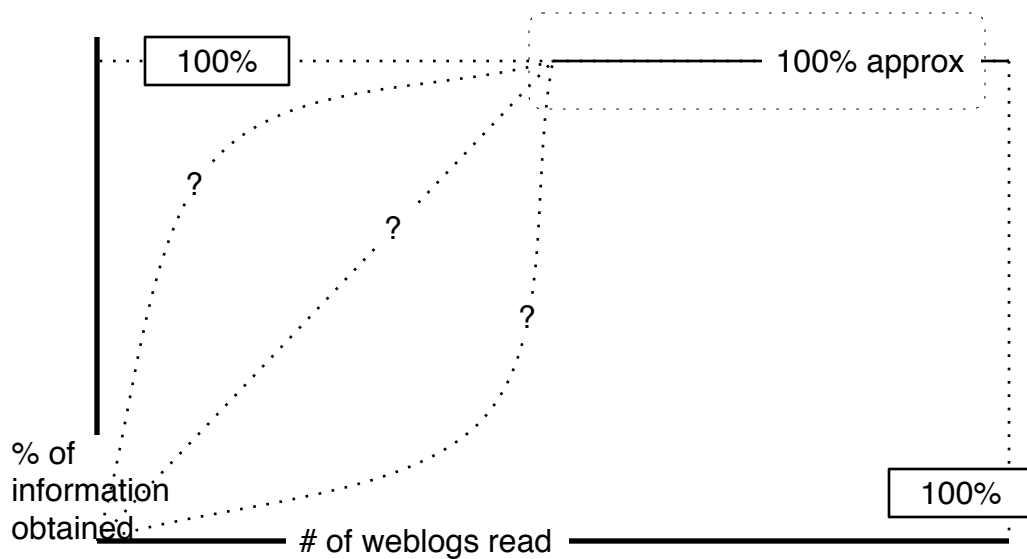


Figure two: Ignore for the moment the dotted lines on the left which represent nothing but the uncertainty of the beginning of the curve. This diagram takes into account that weblogs have different levels of insight within them, and that information is often replicated (either by common, simple insights or by active memetic spread). The 100% threshold will be reached (or be nearly reached) at a much earlier point, with less weblogs read.

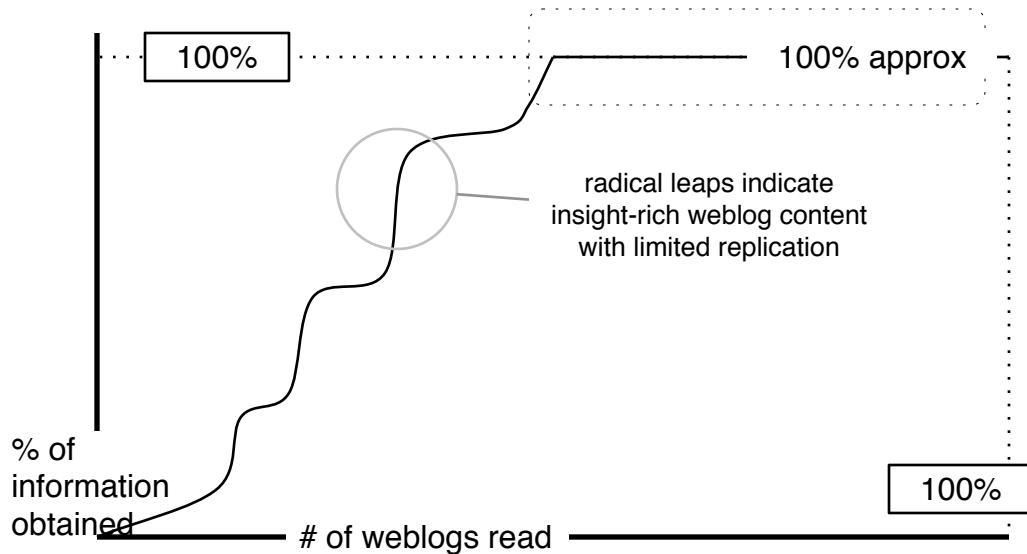
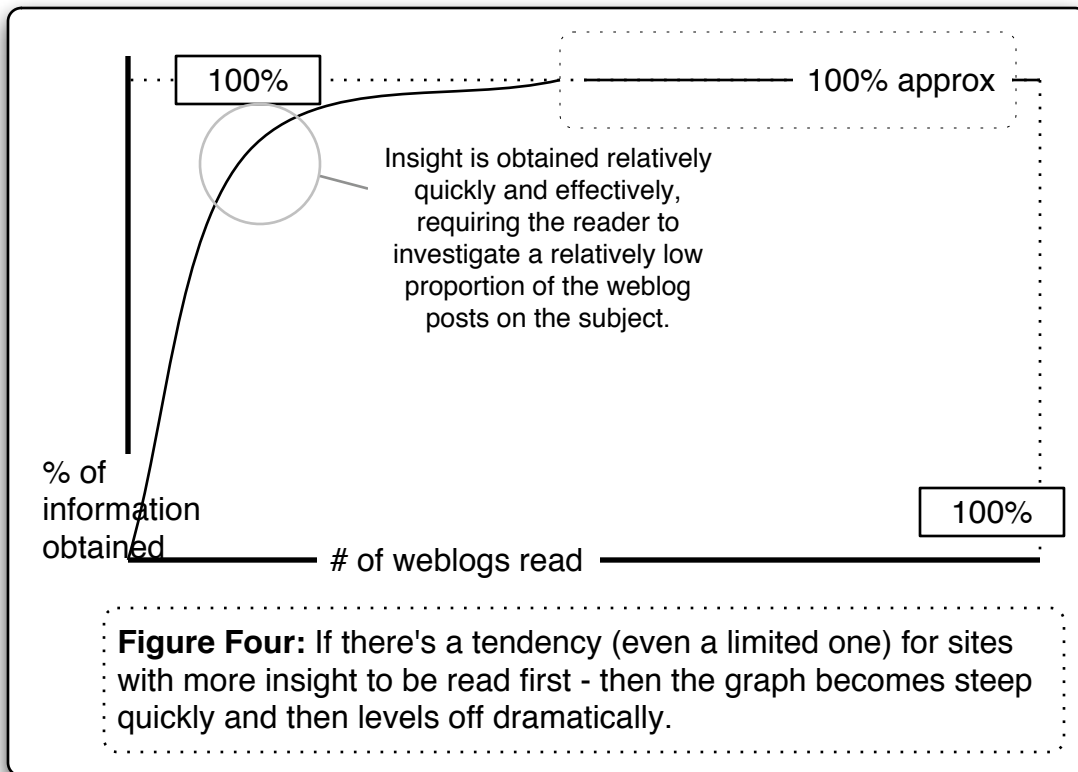


Figure three: In practice - again assuming that we were reading the weblogs in a random order, it would be impossible to gauge the particulars of the curve that lead up to the 100% mark. A sample curve would probably look a bit like this, though - accreting gradually with occasional significant leaps.

Now - all these models have been based upon the assumption that the order in which the weblogs are read will be random. In fact nothing could be further from the truth. Firstly, some weblogs are simply more likely to be read - this is not necessarily purely based upon the level of their contributions, but nor is it completely distinct from such valuations. It would probably be fair to say that well-linked-to sites are more likely (albeit perhaps only incrementally) to contain insight than sites which are not linked to at all. Secondly, if someone does produce content of value and insight, then it is more likely to be linked to - which in turn increases the likelihood that an individual will visit the site in question.

Both of these criteria suggest that (in our attempts to reach the 100% insight threshold) we will be more likely to be directed (initially) to high-insight sites than low-insight sites. This changes our graph substantially.



Hypothetical Conclusions: For any given body of information on weblogs, no matter the rate of replication of information or the number of people who post exactly the same comments, *close to 100% of the available insight* can be reviewed by reading a *disproportionately small number of sites*. **Related Hypotheses:** (1) The *larger* the number of posts about a subject (and hence the more likely replication) the *smaller the proportion* of those sites that need to be read in order to have reviewed close to 100% of the available insight. (2) The size of the available insight will increase as the number of posts about a subject increases (although perhaps not in linear proportion).