Zipf Drive

I'VE LONG been fascinated with the omnipresence of power-law statistics in natural and social phenomena. A good example is Zipf's Law for the usage of English words, named for the 20th-century linguist George Kingsley Zipf. The most common word, *the,* is used twice as often as the second most popular word (*of*) and three times as often as the third (*and*). Similarly, the nth most popular word has a relative frequency of use of $1/n$.
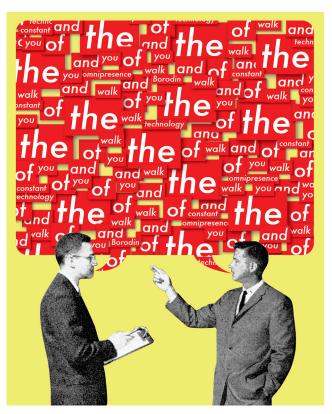
Thus, the curve of popularity versus rank shows a steep decline at first, followed by a long tail that looks rather flat when plotted on a linear scale. (On a log-log plot, of course, this becomes a straight line.) A word like *omnipresence* is way out on the tail, at popularity position 74 228, right before the word *Borodin* (the Russian composer), according to WordCount (http://wordcount.org).

All of the most common words are short, resulting in a very efficient transmission of information. I imagine our distant ancestors sitting around the fire, drawing information-theory equations with sticks in the mud to come up with an optimally parsimonious language, after which they would decide that they shouldn't have used the word *parsimonious* (popularity number 49 309) when something like *concise* would have sufficed.

All this is to say that our vocabulary is rather a perfect blend—100 or so popular words used in everyday conversation and writing, together with about 100 000 more esoteric words that get sprinkled in for effect or special purpose.

Many other phenomena exhibit power-law (that is, polynomial) statistics— cities ranked by population, individuals by wealth, earthquakes by strength, Web sites by number of hits, books by online sales. I would even imagine that it applies to something like the distribution of knowledge in electrical engineering. All of us know Ohm's Law, for example, but perhaps only a tenth of us are familiar with the basic concepts in communications. Then maybe only one engineer in 1000 is familiar with a particular protocol, and only one in 100 000 might be conversant with a particular paper in a specific *IEEE Transactions*. But this is what makes the world go round; we have a lot of things in common, but there is a long tail of specialties that makes each individual unique.

Although power-law statistics have been long known, the subject has gotten much recent attention under the name "the long tail," a phrase coined by Chris Anderson, the editor in chief of *Wired* magazine, in an article in 2004. Discussions have been prompted by the difference between sales in the physical world, where inventories are limited to the popular items, and those in the virtual world of the Internet, where there is no inventory constraint to eliminate all the rare items

on the long tail. In the virtual world, the many small sales out on the long tail approximately equal the sales of the few most popular items.

In most cases there are fundamental reasons that statistics behave like a power law. For example, even though it might seem as if individual choices should be uniformly distributed among alternatives, an individual's choice is often influenced by the choices of others. This explains our herdlike behavior, with a flocking around popular choices and a long tail of individual dissent.

How could it be otherwise? Suppose for a moment that power-law statistics weren't the norm and that choices were uniformly distributed. What would the world be like? With all 100 000 or so words equally likely, books would

be long and turgid but of little interest, because there would be so few subjects of common concern. And of course it would be almost impossible to learn a foreign language.

Population would be uniformly scattered about the Earth. There would be no cities, and whole countries would be like New Jersey, where I have to describe my home's location by the nearest exit number on the Garden State Parkway. For better or for worse, wealth would be uniformly distributed, and perhaps neither cathedrals nor slums would be so prevalent.

I'm sure that you can provide your own suppositions, but perhaps we could all agree that we wouldn't want to inhabit such a world. Our ancient ancestors around the fire figured this out a long time ago. ❑