

6. Power-Law Distributions: Long Tails

Every fundamental law has exceptions. But you still need the law or else all you have is observations that don't make sense. And that's not science. That's just taking notes.

—Geoffrey West

In this chapter, we cover power-law distributions. Often described as long- or heavy-tailed distributions, when graphed these distributions produce a long tail running along the horizontal axis corresponding to large events. The distributions of city populations, species extinctions, the number of links on the World Wide Web, and firm sizes all have long tails, as do the distributions of videos downloaded, books sold, academic citations, war casualties, and floods and earthquakes. In other words, all of these distributions include large events: Tokyo has 33 million residents, J. K. Rowling's Harry Potter books have sold in the neighborhood of half a billion copies, and the great Mississippi flood of 1927 covered an area larger than the state of West Virginia under thirty feet of water.¹

Contemplating a power-law distribution of human heights reveals how much power-law distributions differ from normal distributions. If human heights were distributed by a power law similar to that of city populations, and if we calibrate the mean height at 5 feet 9 inches, then the United States would include one person the height of the Empire State Building, over 10,000 people taller than giraffes, and 180 million people less than 7 inches tall.²


To produce a long-tailed distribution requires non-independence, often in the form of positive feedbacks.³ Book sales, forest fires, and city populations, unlike trips to the grocery store, are not independent. When one person buys a Harry Potter book, she induces others to buy it. When a single tree catches on fire, that fire can spread to neighboring trees. When a city increases in population, it adds amenities and job opportunities, making it more attractive to others. The sociologist Robert Merton referred to the tendency for those who have more to also receive more as the

Matthew effect: “For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken even that which he hath” (Matthew 25:29).

Given the variety of domains in which we find power-law distributions, it would be remarkable if a single mechanism could explain them all, and none does. It would be even more remarkable if each instance of a power-law distribution had a unique explanation. That is also not true. Instead, we possess a collection of distinct models that produce power laws, each capable of explaining different phenomena.

In this chapter, we focus on two models: the preferential attachment model, which explains city sizes, book sales, and web links, and the self-organized criticality model, which explains traffic jams and war deaths, as well as earthquake, fire, and avalanche sizes. In [Chapter 12](#) when we cover entropy, we learn a third model in which a power-law maximizes uncertainty given a fixed mean. And in [Chapter 13](#), we show that return times in a random walk model also satisfy a power law. Still other models show that power laws result from optimal encodings, random stopping rules, and combining distributions.⁴ The remainder of the chapter covers the structure, logic, and functions of power-law distributions, followed by a discussion. The discussion reconsiders the implications of large events and describes the limits of our ability to prevent and plan for them.


Power Laws: Structure



In a *power-law distribution*, the probability of an event is proportional to its size raised to a negative exponent. So for example, the familiar function  describes a power law. In a power-law distribution, the probability of an event is inversely related to its size: the larger the event, the less likely it occurs. Power-law distributions, therefore, have many more small events than large ones.

Power-Law Distributions

A **power-law distribution**⁵ defined over the interval $[x_{min}, \infty)$ can be written as follows:

$$p(x) = Cx^{-a}$$

where the **exponent** $a > 1$ determines the length of the tail, and the constant term  ensures the distribution has a total probability of one.

The size of the power law's exponent determines the likelihood and size of large events. When the exponent equals 2, the probability of an event is proportional to the square of its size. An event of size 100 occurs with probability proportional to , or 1 in 10,000. When the exponent increases to 3, the probability of that same event is proportional to . For exponents of 2 or less, a power-law distribution lacks a well-defined mean. The mean of data drawn from a power-law distribution with an exponent of 1.5 never converges. It increases without limit.

[Figure 6.1](#) shows an approximate graph of the distribution of the number of links to webpages on the World Wide Web.

image

Figure 6.1: Approximate Power-Law Distribution of Webpage Links

The potential for large events distinguishes power-law distributions from normal distributions, from which we practically never see large events. For a long-tailed distribution, though rare, they occur at sufficient frequency to merit attention and preparation. Even one-in-a-million events are worth considering. For example, earthquake sizes approximately satisfy a power law with exponent near two. Suppose that for a region an earthquake larger than size 9.0 on the Richter scale, the size of an earthquake that topples buildings and changes the local topography, occurs each day with a probability of one-in-a-million. Within a century, an earthquake of that size would occur with probability 3.5%.⁶

To see the difference between the probabilities of one-in-a-million events in normal and long-tailed distributions, we can use the distribution of deaths due to terrorist attacks, which follow a power-law distribution with an exponent of 2.⁷ A one-in-a-million event consists of nearly 800 deaths. If deaths due to terrorist attacks followed a normal distribution with mean 20 and a standard deviation of 5, a one-in-a-million event would involve fewer than 50 deaths.

A power-law distribution has a precise definition. Not all long-tailed distributions are power laws. Plotting a distribution on a log-log scale creates a crude test of whether the distribution is a power law. A log-log plot transforms event sizes and their probabilities to their logged values and transforms a power-law distribution into a straight line.⁸



Figure 6.2: Power Law (Black) vs. Lognormal (Gray) on Log-Log Scale

In other words, a straight line on a log-log plot is evidence of a power law, while an initially straight line that gradually falls off is consistent with a lognormal (or an exponential) distribution. The rate at which a lognormal distribution curves downward depends on the variation of the variables that produce the distribution.⁹ As we increase the variance in a lognormal distribution, the tail increases, making it closer to linear on a log-log plot.¹⁰

The special case of power laws with exponents equal to 2 are known as *Zipf distributions*. For power laws with exponents of two, an event's rank times its probability will equal a constant, a regularity known as *Zipf's Law*. Words satisfy Zipf's Law. The most common English word, *the*, occurs 7% of the time. The second most common word, *of*, occurs 3.5% of the time. Notice that its rank, 2, times its frequency of 3.5% equals 7%.¹¹

Zipf's Law

For power-law distributions with an exponent of 2 ($a = 2$), the rank of an event times its size equals a constant.

Event Rank \cdot Event Size = Constant

The populations of cities in many countries, including the United States, are distributed approximately in this way. Using 2016 city population data, each city's rank multiplied by its population produces a value near 8 million.

image

Models That Produce Power Laws: Logic

We now turn to models that produce power laws. Lacking models, power-law distributions remain unexplained patterns.

Our first model, the *preferential attachment model*, assumes entities that grow at rates relative to their proportions. It captures Merton's Matthew effect: more begets more. The model considers a population that grows through arrivals. A new arrival either joins an existing entity or creates a new one. If the latter, the probability of joining an existing entity is proportional to the size of that entity.

Preferential Attachment Model

A sequence of objects (people) arrive one after another. The first arrival creates an entity. Each subsequent arrival applies the following rule: With probability p (small), the arrival forms a new entity. With probability $(1-p)$, the arrival joins an existing entity. The probability of joining a particular entity equals its size divided by the number of arrivals to date.

image

Imagine students coming onto a college campus. The first student creates a new club. With some small probability, the second student creates her own club. More likely, she joins the first student's club. The first ten students might create three clubs: one with seven members, one with two members, and one containing a single member. The eleventh arrival will, with small

probability, create a fourth club. If not, she will join an existing club. When joining an existing club, she chooses the club with seven students 70% of the time, the club with two students 20% of the time, and the club with one student 10% of the time.

The preferential attachment model helps explain why the distributions of links on the World Wide Web, city sizes, firm sizes, book sales, and academic citations are power laws. In each setting, an action (say, a person buys a book) increases the likelihood others will do the same. If the probability of buying from a firm is proportional to its current market share, and if new firms enter at a low rate, then the model predicts that the distribution of firm sizes will be a power law. The same logic applies to book sales, music downloads, and city growth.

Our second model, the *self-organized criticality model*, produces a power-law distribution through a process that builds interdependencies in a system until the system reaches a critical state. A variety of self-organized criticality models exist. The *sand pile model* assumes that someone drops grains of sand onto a table from a spot several feet above. As the grains accumulate, a pile forms. Eventually, the pile attains a *critical state* where additional grains can cause avalanches. At this critical state, additional grains often have no effect or cause at most a few grains to fall. These are the many small events in a power-law distribution. Sometimes the addition of a single grain results in a large avalanche. These are the large events.

A second model, the *forest fire model*, assumes a two-dimensional grid on which trees can grow. Trees can also be hit by random lightning strikes. When the density of the trees is low, any fires caused by lightning will be small, affecting at most a few cells. When the density of trees is high, a fire started by a lightning strike will spread across much of the grid.

Self-Organized Criticality: Forest Fire Model

The forest consists initially of an empty N by N grid. Each period a random site on the grid is chosen. If empty, with probability g the site grows a tree. If the site contains a tree, with probability $(1 - g)$ lightning hits the site. If

the site contains a tree, the tree catches fire, and the fire spreads to all connected sites with trees.

Notice that in the forest fire model, the probability of a lightning strike equals one minus the probability of the growth rate. This construction allows us to vary the relative rate of growth and lightning. It is a simplification that reduces the number of parameters in our model.

Experimenting with the growth rate of trees, we find that for growth rates close to one, the density of trees increases to a critical state: a relatively dense forest of trees, where lightning strikes can wipe out a huge swath of forest. At this critical state, the distribution of the sizes of patches in the forest, and therefore the size of fires, satisfies a power-law distribution. Moreover, the forest naturally tends to this density level. If it is less dense, density increases because fires are small. If density exceeds the threshold, any fire will wipe out the entire forest. Therefore, the tree density *self-organizes* to a critical state.^{[12](#)}

In both the sand pile model and the forest fire model, a macro-level variable—the height of the pile or the density of the forest—has a critical value. That macro-level variable's value decreases when events occur (avalanches and fires). Variants of this model can explain the distributions of solar flares, earthquakes, and traffic jams. An increasing macro-level variable that decreases when events occur, though necessary, is not sufficient for self-organized criticality. Equilibrium systems also have that property. Water flows into and out of lakes through streams, yet because outflows are smooth, lake levels change gradually. The key assumptions for self-organization to critical states is that pressure increases smoothly, like water flowing into the lake, and that pressure decreases in bursts, including possibly large events.

The Implications of Long Tails

We cover three implications of long-tailed distributions: their effect on equity, catastrophes, and volatility. By definition, a long tail means a few big winners (large collapses, earthquakes, fires, and traffic jams) and many losers as compared to a normal distribution, which is symmetric about a

mean. Long-tailed distributions can also contribute to volatility, as random fluctuations in larger entities will have larger effects.

Equity

A person who writes a better book, catchier song, or better academic paper than another should garner more sales and credit. It is not equitable if a person who performs only a little better or who happens to be lucky earns a lot more. As we saw in the preferential attachment model, positive feedbacks create big winners due to the Matthew effect. For positive feedbacks to occur in a market, people must know what others buy, and people must be able to buy the product. For weightless information goods, such as smartphone applications, the latter assumption makes perfect sense. For an iPhone application, no production constraints slow the positive feedbacks as they do for, say, trucks. Ford can only increase production of F-150 trucks by so much. In contrast, Intuit can sell as many copies of TurboTax as people are willing to download.

Empirical studies show that social effects create bigger winners. In the *music lab experiments*, college students could sample and download songs. In the first treatment, subjects did not know what songs others downloaded, and the distributions of downloads had a shorter tail—no song received more than two hundred downloads and only one song received fewer than thirty. In a second treatment, students knew what others downloaded. The tail of the distribution grew: one song received more than three hundred downloads. Perhaps more telling, over half received fewer than thirty. The tail became longer. Social influence increased inequality. This inequality is not a concern if social influence leads people to download better songs. However, correlations between downloads in the two treatments were not strong. If we interpret the number of downloads of a song in the first treatment as a proxy for the song's quality, social influence did not result in people downloading better songs. The big winners were not random, but they were not the best.^{[13](#)}

We must be careful not to draw too strong an inference from a single study. We can, though, infer that while an author who sells 50 million books or an academic whose work receives 200,000 citations deserves accolades, such

extreme success suggests that the central limit theorem is not holding. People are not buying books or making citations independently. Amazing success probably implies positive feedbacks, and perhaps a bit of luck. We return to these ideas when discussing the causes of income inequality in the book's final chapter.^{[14](#)}

Catastrophes

Long-tailed distributions include catastrophic events: earthquakes, fires, financial collapses, and traffic jams. Even though the models cannot predict earthquakes, they provide insight into why their distribution satisfies a power law. That knowledge tells us the likelihood of earthquakes of various sizes. We know what to expect, if not when.^{[15](#)}

The forest fire model does guide action. We can prevent large fires by selectively harvesting trees in a forest to lower the density of trees. Or we might build fire-breaks. One could argue that we do not need a model to tell us to thin a forest or build firebreaks. That is surely true. The model makes us aware that there exists a critical density. That density may vary by forest. It could depend on the type of tree, the prevailing wind speeds, and the topography. The model explains why forests may self-organize to critical states.

We can also use the model as an analogy. Recall that in [Chapter 1](#) we discussed the failures of financial institutions across networks. We can apply the forest fire model to that setting by representing banks and other financial institutions as trees on a checkerboard and allowing adjacencies to correspond to outstanding loans. In that model, a bank failure would be equivalent to a tree catching on fire. That failure could then spread to neighboring banks.

This naive application of the forest fire model to banks would portend large-scale failures as banks become more connected. As we explore that analogy, we see four shortcomings. First, the financial network is not embedded in physical space. Banks can differ in their number of connections. One bank may have dozens of financial obligations while

another may have a mere one or two. Second, trees in a forest cannot take actions to reduce the probability of fire spreading. Banks can. They can increase their level of reserves. Third, the more connected a bank is, the less likely that its failure spreads as its losses will be dispersed across more banks. For example, if a bank defaults on a \$100,000 loan borrowed from a single other bank, that second bank may well go under. If the first bank borrowed the money from a consortium of twenty-five other banks, no single bank takes a large hit. The systems may well absorb the default without collapsing.¹⁶ Last, the spread of a failure from one bank to another depends on the banks' portfolios. If two connected banks hold similar portfolios, then if one fails the other probably is likely already weak. The worst-case scenario occurs if all of the banks in the network hold identical portfolios. In this case, when one bank fails, widespread failure would be likely.¹⁷ If, though, each bank holds a distinct portfolio, poor performance by one need not imply poor performance of another. Bank failures may not spread. A useful model must therefore take into account the assets in the various portfolios. Without this information, knowing which banks have obligations to other banks will be insufficient to predict or prevent failures, and the net effect of greater connectedness of banks will not be clear.

Volatility

Last, we consider a more subtle implication of long-tailed distributions. If the entities that make up a power-law distribution fluctuate in size, then the exponent of the power law becomes a proxy for system-level volatility. It follows that the firm size distribution should influence market volatility. For this exercise, think of a country's gross domestic product (GDP) as the aggregate production of thousands of firms. If production levels are independent and have finite variation, then, by the central limit theorem, the distribution of GDP will have a normal distribution. It also follows that the greater the variation in production levels across firms, the greater the aggregate volatility. If a longer-tailed distribution of firm sizes produces greater variation in production levels, then it will also correlate with greater aggregate volatility.

An examination of volatility patterns in the United States shows that volatility rose in the 1970s and 1980s and then fell for the next two decades in what some call the *Great Moderation*.¹⁸ Beginning around 2000, volatility again increased. It is possible to explain these volatility patterns by changes in the distribution of firm sizes.¹⁹ As the distribution of firm sizes becomes longer- (shorter-)tailed, the largest firms have a disproportionately larger (smaller) effect on volatility. In other words, aggregate volatility increases (decreases) as the firm size distribution becomes longer-(shorter-)tailed. In 1995, when volatility was low, Walmart had revenues of \$90 billion, which corresponded to 1.2% of GDP. By 2016, Walmart's revenues had increased to \$480 billion, or 2.6% of GDP. Walmart's share of GDP more than doubled. In 2016, an increase or decrease in Walmart's revenue would contribute twice as much to aggregate volatility.

No one refutes the logic of this argument. The relevant question becomes whether a calibrated model produces effects with magnitudes that correspond to actual volatility levels. The calibrated fit proves quite close. Firm size distributions correlate nicely with the historical evidence of the Great Moderation. That correlation does not prove that it is changes in firm size distribution (instead of effective government management of the economy or better inventory control) that caused the moderation, but it does prevent us from rejecting the model.²⁰ The evidence also provides reason to keep this model in our quiver when we evaluate fluctuations in the future.

Contemplating a Long-Tailed World

In long-tailed distributions, large events occur with sufficient probability to be of concern. In the models we covered, long-tailed distributions arise because of feedbacks and interdependencies. We should pay heed to that observation. As our world becomes more interconnected and feedbacks increase, we should see more long tails. And the current long tails that we see may get stretched even further. Inequities may increase, catastrophes grow larger, and volatility become more pronounced. None of these is desirable.

So far, we have discussed these possibilities at macro levels. They also occur at smaller scales. Boston's "Big Dig," a three-and-a-half-mile tunnel through the center of the city, provides an example of a moderate-scale catastrophe. The project cost taxpayers \$14 billion, more than three times the original estimate, and it became the most expensive highway project in the history of the United States. Model thinking frames the Big Dig not as a single project but as an aggregate of subprojects: digging a trench, pouring a concrete tunnel, engineering a drainage system, and building walls and a roof. The project's total cost equals the sum of the subprojects' costs.

If the costs of each subproject had been additive, then the distribution of costs for the project would have been normally distributed.²¹ However, the subprojects' costs were connected. When the epoxy used to glue the roof into place proved inadequate, it was replaced with a costlier, stronger epoxy and, therefore, raised the cost of the project. The failure of the first epoxy created additional costs associated with removing and replacing the collapsed roof. Those efforts in turn required redoing several other parts of the project. Overall costs more than doubled because each project had to be undone and then redone. Interdependencies led to a large, and costly, event.

The potential for large events makes planning difficult. The distribution of natural disasters such as earthquakes satisfy a power law. Thus, most events will be small, but some will be large. If catastrophic events follow a power-law distribution with an exponent near 2, then governments need to keep a very large amount of money in reserve or at least at the ready. They need to prepare for a very rainy day. If governments do so by maintaining huge surpluses in an emergency fund, they may be able to stop themselves from spending that money or cutting taxes if no large event occurs.

Search and Opportunity

We can apply our knowledge of distributions within a class of search models to explain why the number of opportunities a person receives may correlate strongly with success. We embed one class of models, our distribution models, within a second class, search models. When we search,

whether for a new pair of shoes, a career, or a vacation spot, we do not know our choice's value until we try it, though we may know something about the distribution of values, such as the mean, standard deviation, and whether the distribution is normal or has a long tail.

Here, we model choice of profession as a search process. Given a profession, a person tries a career path, which we model as a draw from a distribution. We assume that she can either stick with that career or try again. Trying again corresponds to another draw from the distribution. Consider, for example, the choice of profession for a talented young scientist. She could go to medical school or do research in quantum computing. Medical school offers the safer path. Choosing to work on quantum computing involves becoming an entrepreneur and taking on more risk. To account for these differences, we represent the salary distribution for doctors as a normal distribution with mean \$250,000 and a standard deviation of \$25,000, and the salary distribution for the entrepreneurial career as a power law with an exponent of 3 with an expected salary of \$200,000.^{[22](#)}

Within each profession, our scientist can try multiple careers. She can search. A doctor can switch from oncology to radiology. A failed entrepreneur can pick up the pieces from her start-up and try anew. Each career switch entails a cost. For a doctor, it means more training. For an entrepreneur in quantum computing, it means more long nights of work with little to no compensation.

We assume that our young scientist finds the two professions equally stimulating and makes her choice based on salary. Our model reveals that the better choice depends on how many times she can afford to try new careers. If she must stick with her first career choice, becoming a doctor offers the higher expected salary. If she has sufficient resources to continue trying to be an entrepreneur, eventually she will get a high-paying draw from the long tail. The figure below shows the average largest salary across twenty trials assuming one, two, five, and ten career searches within each profession. If she has the opportunity to try her hand at quantum computing start-ups ten times, her salary will be nearly double what she would earn had she chosen medical school and experimented with ten careers.



Average Income as a Function of Number of Opportunities

If wealth and family support correlate with the number of opportunities a person has to try new careers, our model predicts that wealthier people will choose riskier professions.²³ Evidence on patents aligns with the model. The probability that someone writes a patent correlates with that person's mathematical abilities. People in the top 1% of math ability are far more likely to hold a patent. Among the top 1%, those from families in the top 10% of the income distribution are even more likely to hold a patent.²⁴ At least two models could explain the disparity. One model could assume that poorer talented students never attend college. They may be working routine jobs and never have the choice between medical school or quantum computing. Or, perhaps poorer students choose safer careers.

The logic that an increase in opportunities creates an incentive for risk applies widely. Venture capitalists take risks because they make multiple investments. An early investment in a single unicorn, a billion-dollar company, more than compensates for many losers. Pharmaceutical research laboratories also take risks, spending billions on drug research. We can apply the same logic when deciding where to eat lunch. When driving cross-country and stopping in an unfamiliar town, we may want to eat at a chain restaurant. If moving to that town, we should experiment.