POWER LAWS AND POLYA URNS

Random graphs and dependence

A random graph with n nodes is created by the following process: For each unordered pair of nodes [i,j] the edge [i,j] is present with probability d/n, *independently* of the other edges. From the point of view of a node, the edges adjacent to this node are distributed according to a binomial distribution with mean d and variance approximately dn (assuming that n >> d).

We know that the binomial distribution (like all distributions we have seen so far) is very concentrated about the mean. For example, if a graph with n = 20,000 nodes has average degree d = 4, we do not expect any node to have degree 20 or more. But in the Internet, the graph of the 20,000 or so autonomous systems which indeed has average degree near 4, *about 40 nodes have degree greater than 20!* And in the www (the graph of hyperlinks between n = 10,000,000 documents, if we ignore directions), a graph with average degree about 12, *more than 100,000 have degree at least 100!* We must conclude that there is some kind of dependence between the edges of these graphs, that they came about by a process that is very different from independent coin tosses. The degrees of these graphs seem to have a density function that does not decrease exponentially with |x – m| (the difference from the mean)…

Power Laws

The largest city of the USA has population pop_1 = 15,000,000, the second-largest, pop_2 = 8,000,000, the third largest perhaps pop_3 = 4,000,000, and so on. If you plot $\log(pop_i)$ as a function of log(i), you will notice that it is almost a straight line with angle 45%:

$$pop_i \approx pop_1 / i$$

Similarly, if $freq_i$ is the frequency of the i-th most frequent word in a corpus (say, Shakespeare's plays), then

$$freq_i \approx freq_1 / i.$$

This phenomenon is quite pervasive, and it is called *Zipf's Law*.

There is a similar phenomenon in the distribution of incomes, first observed by the economist Pareto. He noticed that

$$\text{prob}(\text{income} > x) = D\, x^{1-a}$$

The density function of this distribution (obtained by differentiating the rhs) would be $C\, x^{-a}$ Notice that it decreases *polynomially*, not exponentially, with x.

To compare this with Zipf's Law, let's see how much is the income of the i-th richest person. If x is the income of the i-th richest person, then, if we have a population of N people, the number of people with income ≥ x will be

$$i = ND\, x^{1-a}$$

Solving this for x = inc_i (the i-th largest income) we get

$$inc_i \approx inc_1 / i^{1/(a-1)}$$

(What is the formula for inc_1 in terms of N, D, and a?) This kind of distribution is called a *power law*. Comparing with Zipf's law, we see that a power law is more general: Zipf's law is a power law with a = 2. If we plot the power law $D\, x^{1-a}$ or its density function we see that they go to zero much slower than the normal or binomial distributions (which go to zero exponentially fast); that is why they are often called "heavy tailed" distributions.

We know that the exponentially concentrated distributions are produced by many independent contributions (as in the binomial distribution, or the normal distribution, viz. the law of large numbers). What kind of processes produce power law distributions? There are several processes that are known to produce such distributions:

- ***Size-independent growth:*** If the ratio of a person's income next year to this year's income is a random variable that *does not depend on the income*, then it can be shown that, after a while, a population whose incomes evolve this way will obey a power law. This is how a network evolves: high-degree nodes are "important," and attract more edges.
- ***Trade-offs:*** If people try to optimize a single objective, concentrated distributions often result. But if they must optimize – trade one against the other – two or more objectives simultaneously, then the optimum usually has power law-distributed features. Entities interacting in a complicated environment like the Internet and the www are likely to behave this way.
- ***Copying:*** The web is arguably constructed by new documents creating hyperlinks to documents discovered by following other hyperlinks. Copying results in a power law distribution.
- ***Preferential attachment*** For a slightly different idea, in the web a document with many hyperlinks is considered more important, and a new document is more likely to be linked to it. This results in power law.

Polya Urns

There is a simple model that captures the essence of preferential attachment. Suppose that you have two bins, the white and the black bin, with w and b balls inside, respectively, and you throw a new ball. The balls are magnetic, and they attract the new ball with force proportional to their number; so the new ball will go to the white bin with probability w/(w+b). Suppose we start with w = b = 1, and throw in n new balls. How is the resulting distribution going to look like? This process is called a *Polya urn*. Intuitively, the "attraction" rule is going to enhance inequalities, so the bins are not going to be very load-balanced. But how exactly?

For example, what is the probability that, if we through in 1,000 balls, the split is going to be 90% to 10% in favor of black, or worse? The answer is simple: ***10%.*** (Notice that, in ordinary balls and bins, a 90-10 split is extremely unlikely.)

Here is why. Suppose that I generate a permutation of {0,1, 2, …, n} as follows: I start with a ball labeled 0 somewhere on the line. Then I through in a ball labeled 1; it will end up left or right of 0 with probability 50%. Suppose it ends up to the right of 0. Then I throw ball 2. It will end up with equal probability to the left of 0, to the right of 1, or between 0 and 1. When I throw ball 3, there are 4 possible positions. And so on. It is easy to convince yourself that, this way, all permutations are generated with equal probability.

Now suppose that I call the balls that ended up to the right of ball 0 ``white'' and the ones to the left "black". It is easy to see that at each step the new ball is white with probability exactly w/(w+b); hence this is precisely the Polya urn described above!

One more observation: Where is 0 going to be? Since this is a random permutation, 0 can be at any one of the n + 1 positions with equal probability. Hence, the number of (new) white balls is going to be 0,…,n with equal probability! And the probability that the 0 ball will end up in the 10 percentile is exactly 10%...

Suppose now that you have n (new) balls and k colors, not just two. This is the same as having n + k -1 balls, where the first k – 1 ones are just "color boundaries". And the end result is tantamount to selecting k – 1 from among these balls: The first color is just the balls to the left of the first boundary, the second consists of the balls between the first and second boundary, and so on. How does this compare to balls and bins?

The maximum will be about (n/k) log n, only a log n factor above the expectation. But the minimum will be much smaller: n/k^2. In this sense, Polya urns foster imbalances.

There is a model of the web that is a slight extension of Polya urns: Suppose that, once a new ball is thrown, it is added to one of the bins, *but also a new bin is started, with one ball in it.* This models a graph where nodes arrive, and upon arrival each node u has one edge directed out of it; this edge is directed to a previous node v with probability equal to the degree of the node v at the time. The in-degrees of such graphs are known to be power law-distributed with a = 3.