1 Degree distributions and data

A great deal of effort is often spent trying to identify what functional form best describes the degree distribution of a network, particularly the upper tail of that distribution. The most popular model for degree distributions is, by a large margin, the power law, but alternatives include the log-normal, the stretched exponential (also called the Weibull), a power-law distribution with an exponential cutoff in the upper tail, and the exponential.

The table below gives the mathematical definitions of many distributions. It is important to remember that degrees are discrete variables, and their distribution is described by a discrete function. That said, however, when $x_{\min}$ is very large, the difference between a continuous and a discrete distribution with the same functional form is often negligible, meaning that we can approximate the discrete distribution with the continuous form.

	name	distribution $p(x) = Cf(x)$	
		$f(x)$	C
continuous	power law	$x^{-\alpha}$	$(\alpha-1)x_{\min}^{\alpha-1}$
	power law with cutoff	$x^{-\alpha}\mathrm{e}^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{\min})}$
	exponential	$\mathrm{e}^{-\lambda x}$	$\lambda\mathrm{e}^{\lambda x_{\min}}$
	stretched exponential	$x^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$	$\beta\lambda\mathrm{e}^{\lambda x_{\min}^{\beta}}$
	log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\mathrm{erfc}\left(\frac{\ln x_{\min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$
discrete	power law	$x^{-\alpha}$	$1/\zeta(\alpha,x_{\min})$
	Yule distribution	$\frac{\Gamma(x)}{\Gamma(x+\alpha)}$	$(\alpha-1)\frac{\Gamma(x_{\min}+\alpha-1)}{\Gamma(x_{\min})}$
	exponential	$\mathrm{e}^{-\lambda x}$	$(1-\mathrm{e}^{-\lambda})\,\mathrm{e}^{\lambda x_{\min}}$
	Poisson	$\mu^x/x!$	$\left[\mathrm{e}^{\mu}-\sum_{k=0}^{x_{\min}-1}\frac{\mu^k}{k!}\right]^{-1}$

Table 1: Definition of the power-law distribution and several other common statistical distributions, many of which are proposed as models of degree distributions in networks. For each distribution we give the basic functional form $f(x)$ and the appropriate normalization constant C such that $\int_{x_{\min}}^{\infty} Cf(x)\,\mathrm{d}x = 1$ for the continuous case or $\sum_{x=x_{\min}}^{\infty} Cf(x) = 1$ for the discrete case.

2 Power-law distributions and data

Suppose we have some empirical observations $\{x_i\} = \{x_1, x_2, \ldots, x_n\}$ to which we would like to fit a power-law distribution. Recall from the previous lecture that there are two parameters we need to know to do this: α, the "scaling" exponent, and $x_{\min}$, the smallest value for which the power law holds.

2.1 Estimating α

For the moment, let us assume that the correct value of $x_{\min}$ is known. To choose a good value of α, we apply the principle of maximum likelihood, which chooses the parameter (often generically denoted θ) that maximizes the likelihood of observing exactly the data $\{x_i\}$ under the model.[1] This is done by writing down a function for the *likelihood*, which depends on the data, the model, the model's parameters.

$$\mathcal{L}(\{x_i\} \,|\, \theta) = \prod_{i=1}^{n} p(x_i \,|\, \theta) \tag{1}$$

We then want to find the value of θ that maximizes this function, with respect to the model and data. Note that one can always carry out this calculation, given a choice of a model, but the result of this exercise gives you no information about whether the model is a *good* fit to the data.

For complex models, it may not be possible to derive a closed-form expression for α in terms of $\{x_i\}$. However, the power law is not a complex model. Thus, substituting the normalized form of the power law's $p(x)$ into Eq. (1), we have

$$\begin{aligned}\ln \mathcal{L}(\{x_i\} \,|\, \alpha, x_{\min}) &= \ln \left[\prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha} \right] \\ &= n \ln \left(\frac{\alpha - 1}{x_{\min}} \right) - \alpha \sum_{i=1}^{n} \ln \left(\frac{x_i}{x_{\min}} \right) \quad .\end{aligned}$$

Taking the logarithm of the likelihood, the *log-likelihood*, is a standard trick that often makes the function easier to work with by replacing product series with summations and allowing computers to more easily represent ridiculously tiny values like 10^{-3000}. More importantly, the parameter that maximizes the log-likelihood is the same as the parameter that maximizes the likelihood. (Why?)

For more complex models, the next step is often to numerically maximize this function, perhaps using non-parametric techniques like Nelder-Mead or hill-climbing techniques like gradient ascent.

[1] There are generally very good reasons to choose parameters in this way, in part because the maximum likelihood choice is a *consistent* estimator, meaning that $\hat{\theta} \to \theta$ in the limit of $n \to \infty$, and its errors are asymptotically normal.

However, the power law is not a complex model, and we may derive an analytic expression for the location of its maximum. Solving $\partial\mathcal{L}/\partial\alpha = 0$ for α yields

$$\hat{\alpha} = 1 + n \Bigg/ \sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min}}\right) \quad , \tag{2}$$

which is the MLE or maximum likelihood estimator for α. (We use "hatted" variables to denote estimates derived from data, and unhatted variables to denote the true, but generally unknown, value.) Via similar process,[2] we may also derive a closed-form expression for an estimator of the standard error in $\hat{\alpha}$, which is $\hat{\sigma} = (\hat{\alpha}-1)/\sqrt{n} + O(1/n)$.

A crucial assumption of this process was that the value of $x_{\min}$ was already known, but in general we do not know this value. Why is this a crucial assumption? Because $x_{\min}$ is not a normal parameter and instead truncates or "censors" the data by changing the number of observations that go into the likelihood function. (What would the maximum likelihood choice of $x_{\min}$ be, and why is this a useless result?) Thus, we need a different method by which to choose $x_{\min}$.

2.2 Estimating $x_{\min}$

There are a number of methods by which to choose $x_{\min}$, the most common of which is to choose it subjectively, by eye. But, this is clearly not a good method.

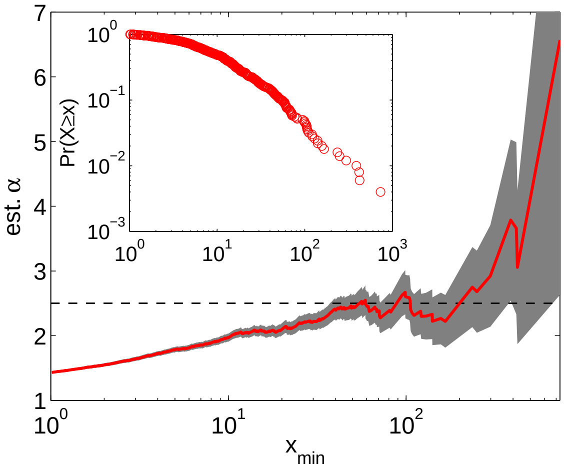
2.2.1 The Hill plot

For the power-law distribution in particular, a common technique for choosing $x_{\min}$ is a visual diagnostic called a "Hill plot," which plots $\hat{\alpha}$ as a function of $x_{\min}$. In the power-law region of the distribution, the value of α should not increase or decrease even if we raise $x_{\min}$, and thus we should choose the smallest value of $x_{\min}$ that produces such a stable estimate of α. This is generally done visually.

This approach, however, is not reliable. For instance, the figure below shows a Hill plot for $n = 500$ observations drawn iid from a shifted power law $p(x) \sim (k + x)^{-\alpha}$, with $k = 15$ and $\alpha = 2.5$. The inset shows the ccdf of the actual data, and the shaded region shows the uncertainty around $\hat{\alpha}$.[3] Visually, things start getting flat somewhere around $x_{\min} \approx 20$, but this yields $\hat{\alpha} \approx 2$, which is a much heavier tail than is accurate. The deviations from the true value of α for smaller values of $x_{\min}$ are caused by fitting a power-law model to non-power-law data. The deviations for larger values of $x_{\min}$ are caused by variance induced by sampling noise and by a small-sample bias in

[2] The standard error is given by the curvature of the likelihood function at the location of the maximum, which is related to the Fisher information of the function.

[3] Recall from the last lecture that for a shifted power law, there is no *correct* choice of $x_{\min}$, above which the distribution follows a pure power law. Thus, we simply want a value of $x_{\min}$ that yields an accurate estimate of α.



the MLE for α. For these data, there is no visually obvious choice of $x_{\min}$ that yields an accurate estimate of α because these two regions overlap significantly.

An important message: In the wide world of data analysis techniques, there are many visual diagnostic tools with a similar flavor. Be forewarned: keep your statistical wits about you when you encounter them. *Looking* at your data is a very important part of data analysis because there is no better way to get an unfiltered view of your data than through simple visualizations: plot distributions of quantities, scatter plots of attributes, time series, etc. The trouble enters when visualization requires the use of statistical tools that have built-in assumptions, and all data analysis tools have built-in assumptions. Thus, here is an important take-home message: *in order to properly understand your data, you must first properly understand your statistical tools*—what input they require, what input they *expect*, what operations they perform on the data, why they performs those and not other operations, and finally how to interpret their output correctly.

The Hill plot diagnostic fails to provide a *reliable* way to choose $x_{\min}$ because it does not provide a quantitative and objective answer to the following questions: (i) How do we automatically quantify "flat"-ness? And, (ii) given a method to do so,[4] how "flat" is flat enough?

[4] For instance, we could fit a first-order polynomial to the $\hat{\alpha}(x_{\min})$ function above some choice of $\hat{x}_{\min}$.

2.2.2 KS minimization

An objective and fairly accurate solution to choosing $x_{\min}$ is to minimize a so-called "goodness-of-fit" (gof) statistic,[5] such as the Kolmogorov-Smirnov (KS) statistic, between the fitted model and the data. The KS statistic is defined as

$$D = \max_{x \in \{x_i\}} \left| P(x \,|\, \hat{\theta}) - S(x) \right| \quad , \tag{3}$$

where $P(x \,|\, \hat{\theta})$ is the theoretical cdf, with parameters $\hat{\theta}$ and $S(x)$ is the empirical cdf.[6]

The KS statistic (typically denoted D) measures the largest deviation in cumulative density between the fitted model and the empirical data. Figure 6 shows an example of this for a small sample from an exponential distribution. Because D measures the maximum deviation, a small D implies that the fitted model $P(x \,|\, \hat{\theta})$ is everywhere close to the empirical distribution. (What kind of deviations is the KS statistic most sensitive to? For what kind of questions might this behavior matter?)

Suppose that we choose $x_{\min}$ too small, such that we fit the power-law model to the distribution's "body" where there are significant deviations from the power-law behavior. In this case, D should be large because of a model-misspecification bias coming from the data. On the other hand, if we choose $x_{\min}$ too large, and fit the model to the distribution's extreme tail where there are few observations, statistical noise or variance in the data will make D large. We want to choose an $x_{\min}$ somewhere between these two, that is, we want a balanced tradeoff between bias on the one hand and variance on the other. This can be done by choosing $\hat{x}_{\min}$ in the following way:

$$\hat{x}_{\min} = \inf_{x_{\min} \in \{x_i\}} \left(\max_{x \geq x_{\min}} |P(x \,|\, \hat{\alpha}, \hat{x}_{\min}) - S(x)| \right) \quad . \tag{4}$$

That is, we estimate $\hat{x}_{\min}$ as the value of x_i that yields the smallest maximum deviation. Note that each time we increase our candidate value for $\hat{x}_{\min}$, we need to truncate the data set, re-estimate α and compute a new theoretical cdf and empirical cdf.

Using the same data in the figure above, we can apply the KS minimization technique, whose results are shown in the figure below, with the KS-minimizing value of $x_{\min}$ marked, and the resulting power-law tail fit shown in the inset with the data (ccdf). The result is $\hat{x}_{\min} = 39.79$; at this choice, we get $\hat{\alpha} = 2.41 \pm 0.16$, which is indistinguishable from the true value of $\alpha = 2.5$.

[5] There are a number of other such statistical measures, including the sum-of-squared errors, the weighted-KS statistic, etc. These statistics generally have nice mathematical properties, which is why there are commonly used. In general, all such measures quantify the magnitude and direction of deviations between the observed data and some model of that data.

[6] The empirical cdf $S(x)$ is a step function, defined as the fraction of the full data set that are less than some value x. If we sort our data so that $x_1 < x_2 < \cdots < x_n$, then the corresponding y values for the empirical cdf, in order, are $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\}$.

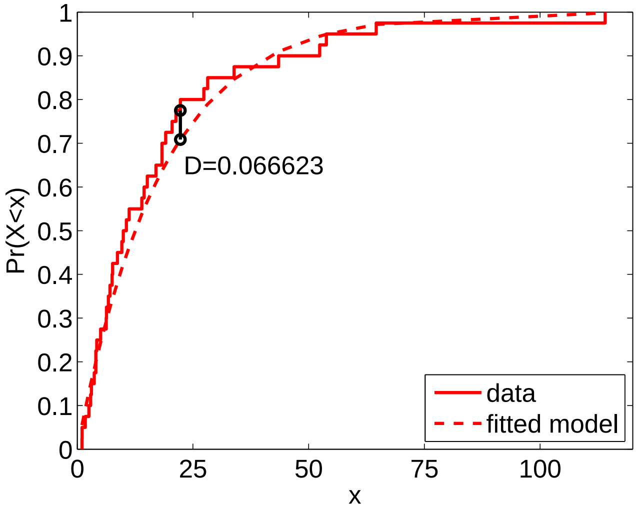
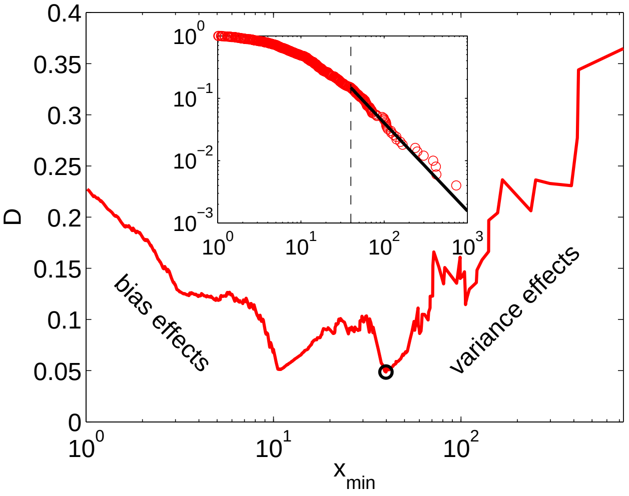


Figure 1: The empirical cdf (solid line; $n = 40$, $\lambda = 0.050$) and the maximum likelihood theoretical cdf (dashed line; $\hat{\lambda} = 0.053 \pm 0.009$) for an exponential distribution. The black line shows the maximum absolute deviation between the two functions of $D = 0.066623$.

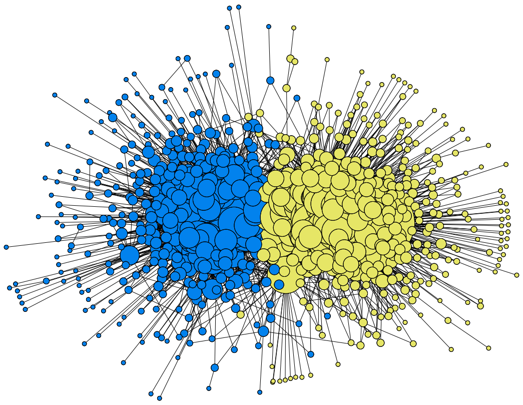
This agreement between $\hat{\alpha}$ and α, however, is slightly misleading. Recall that these data are drawn from a shifted power law with $k = 15$; at $\hat{x}_{\min} \approx 30 = 2k$, there should still be non-trivial deviations from the pure power-law model we're fitting (see previous lecture). In fact, I had to generate several synthetic data sets to get a fit this good. The point here is that KS minimization provides an automatic, objective and fairly reliable way to choose $x_{\min}$, even if it doesn't provide the strong guarantees we expect from a maximum likelihood procedure (nice things like asymptotic consistency and normally-distributed errors). That is, this is a nice example of a reasonable computational procedure for solving a tricky statistical problem.

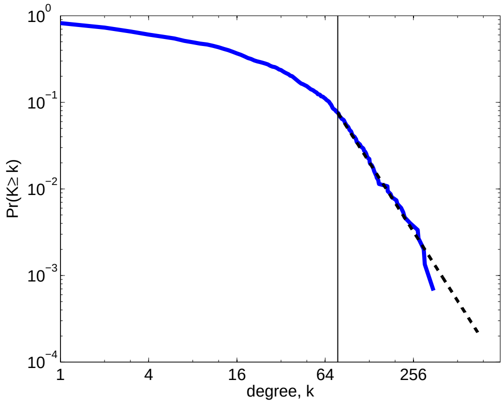
Numerical experiments on its accuracy show that it is also quite reliable, although it does make some mistakes. It is good that we can measure and quantify these mistakes on synthetic data, where we know what the true underlying structure is: it means we can learn how to interpret its behavior, including potential mistakes, when we apply it to empirical data with unknown structure. In general, these numerical experiments suggest that when applied to data drawn from a distribution that actually exhibits a pure power-law form above an explicit value of $x_{\min}$, KS minimization is slightly conservative, i.e., it tends to choose an $\hat{x}_{\min} \gtrsim x_{\min}$. Unfortunately, it is not known in general how large a bias KS minimization can exhibit or even why, mathematically, it works so well in practice.



2.3 Degree distribution of the political blogs network

Recall from the last lecture that the political blogs network has a heavy-tailed degree distribution (see figure below). Taking the KS-minimization approach for choosing $x_{\min}$ and using the discrete power-law distribution as the model (see Table 1 above), we may estimate the best-fitting power-law tail model for these data. The resulting fit chooses a portion of the upper tail that visually looks linear on the log-log plot, for degrees $k \geq 78$ (which is 113 vertices, or about 7.6% of the network), and yields $\hat{\alpha} = 3.65$. The right-hand panel of the below figure shows the result, with a vertical line drawn at $\hat{x}_{\min}$.





These numbers for the fitted power-law model do not shed any light on whether the model is a good fit to the data. To answer that question, other tools are needed. For example, we could use a statistical hypothesis test to determine whether drawing values iid from the fitted model is a plausible explanation for the values or a likelihood ratio test to determine whether this model is a better fit than some alternative model.

3 At home

1. Read Chapter 8.1–8.4 (pages 235–260) in *Networks*
2. Next time: social and biological networks