Internet Mathematics Vol. 6, No. 1: 19–28

An Occupancy Problem Arising in Power Law Fitting

Ian Abramson and Arthur Berg

Abstract. The power law arises commonly in modeling the number of vertices of a given degree in large graphs. In estimating the degree of the power law, the typical approach is to truncate by eye the log-log plot, then fit a linear equation to the remaining log-transformed data. Here we formulate a hard-coded truncation rule to replace the visual truncation, justify it by showing that the truncation point goes to infinity and misses a vanishing fraction of the data with probability tending to one, and refine the subsequent regression with a weighting and a way to use the covariation between slope and intercept to optimize the slope estimate.

1. Introduction

The power law is widely used to model the number of edges at the vertices of large graphs. The emergence of scale-free networks in recent years that model real networks has the defining characteristic that the probability of a node connecting to k other nodes follows a power law distribution [Barabási and Bonabeau 03, Ravasz and Barabási 03]. The suitability of this heavy-tailed distribution in preference to other close shapes is an empirical matter, not strongly supported by any mechanistic rationale, and popular in part because a certain plot of transformed count and frequency data leads transparently to a fitting method based on simple regression.

1542-7951/08 $0.50 per page

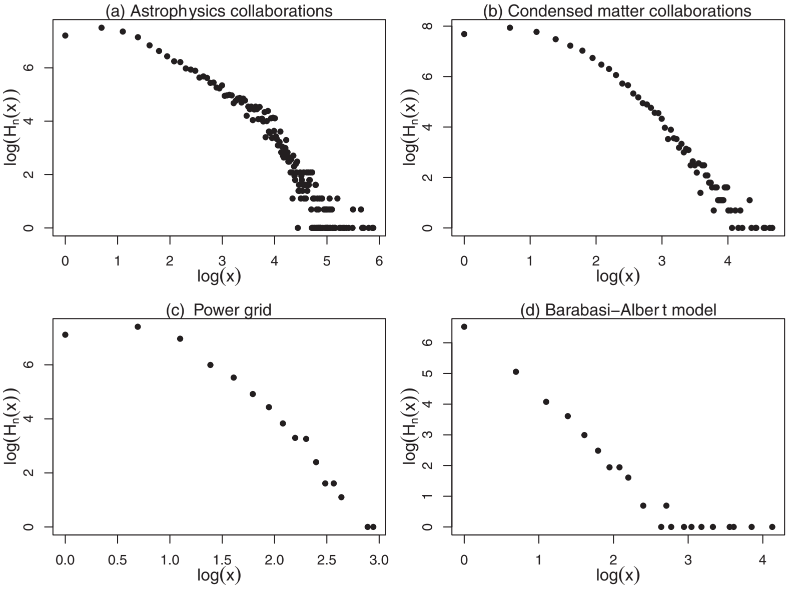

Figure 1. Example log-log plots from real and simulated networks demonstrating the frequently observed "hook" on the left, the need for truncation on the right, and the need for weighting the regression. (a) Network of coauthors from the Astrophysics E-print Archive [Newman 01]; (b) network of coauthors from the Condensed Matter E-print Archive [Newman 01]; (c) US Western States Power Grid network [Watts and Strogatz 98]; and (d) Barabasi–Albert scale-free graph simulation [Barabási and Albert 99] with 1000 nodes and 1998 vertices.

The idealized data model in a large random sample $X_1, \dots, X_n$ forms a discrete power law distribution

$$f(x) = f_\beta(x) = \frac{c}{x^\beta}, \quad x = 1, 2, \dots,$$

for some $\beta > 1$; note that $c = c_\beta = \frac{1}{\zeta(\beta)}$, where ζ is the Riemann zeta function. Typically there are unknown weak local dependencies in real large-graph data that the model ignores, along with possible demonstrable lack of fit for small x. The fitting requirement here is to capture the tail behavior in estimating β in some way that is robust against such departures from the ideal model.

We shall refer to the log-plot as

$$\{(\log x, \log H_n(x)) : x = 1, 2, \dots\},$$

where $H_n(x)$ is the count of vertices of degree x, i.e.,

$$H_n(x) = n\hat{f}(x) = \sum_{i=1}^{n} 1[X_i = x].$$

Logs of zero counts are not plottable. Low counts and zero counts are expected on the right, where the linear signal is blurred by increasing noise and where the points are most influential on the regression. The unplottable zero counts for large x induce a bias for estimating $\log nf(x)$, and inclusion in the regression is undesirable. Our chief purpose here is to develop a right truncation point for the points included in the regression so that the difficulty arises with probability tending to 0. (In practice, the plot often shows a “hook” on the left, cf. Figure 1, an empirical phenomenon with an ad hoc solution: delete by eye!) The regression will be weighted, with weighting scheme guided by a delta-method calculation. The deterministic sequence y_n of truncation points given in the main theorem is determined up to a growth condition, but not fully specified. In reality, a user will usually truncate by eye anyway, and perhaps the main use of the theorem lies in guaranteeing that a nonwasteful truncation will be possible at about the end of the usable linear region.

2. Weighting of the Regression and Optimal Use of the Fitted Coefficients

The weighting scheme requires knowing the variance of $H_n(x)$, the frequency histogram at a fixed value of x. Since $H_n(x)$ is a binomial random variable with parameters n and $cx^{-\beta}$, the central limit theorem implies

$$\frac{H_n(x) - ncx^{-\beta}}{\sqrt{ncx^{-\beta}(1 - cx^{-\beta})}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1).$$

After invoking the delta method, we obtain the following convergence:

$$\sqrt{n}\left(\log H_n(x) - \log(ncx^{-\beta})\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, c^{-1}x^{\beta} - 1).$$

Thus the asymptotic variance depends on β.

Strictly, this would call for iterative reweighting. At the initial stage, one would truncate the plot by eye and perform an unweighted or crudely weighted regression to get a preliminary estimate of β from the slope and c from the intercept. Now reweight the points inversely to $(x^{\beta} - c)$ and iterate, moving the right truncation point, which is also guided by β according to the main theorem, which follows. Note that since c depends on β as well, an optimal estimate of β

from the regression approach would be a convex combination of $\hat{\beta}$ and a mapping of $\hat{c}$.

Specifically, note that

$$\beta = \beta(c) = \zeta^{-1}(1/c) \quad \text{and} \quad \beta'(c) = \frac{-1}{c^2 \zeta'\left(\zeta^{-1}(1/c)\right)}.$$

The weighted regression statistics include an estimate $\left(\begin{smallmatrix} \hat{\tau}_{11} & \hat{\tau}_{12} \\ \hat{\tau}_{21} & \hat{\tau}_{22} \end{smallmatrix}\right)$ for the covariance of $\left(\begin{smallmatrix} \hat{\beta} \\ \hat{c} \end{smallmatrix}\right)$, the fitted weighted least-squares coefficients. Basing another estimate of β on $\hat{c}$ given by $\tilde{\beta} = \zeta^{-1}(1/\hat{c})$, we have a correlated pair $\left(\begin{smallmatrix} \hat{\beta} \\ \tilde{\beta} \end{smallmatrix}\right)$ of asymptotically unbiased estimates of β with covariance efficiently estimated by

$$\begin{pmatrix} \hat{\tau}_{11} & \tilde{\tau}_{12} \\ \tilde{\tau}_{21} & \tilde{\tau}_{22} \end{pmatrix} := \begin{pmatrix} 1 & 0 \\ 0 & \beta'(\hat{c}) \end{pmatrix} \begin{pmatrix} \hat{\tau}_{11} & \hat{\tau}_{12} \\ \hat{\tau}_{21} & \hat{\tau}_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \beta'(\hat{c}) \end{pmatrix}.$$

There is an optimal convex combination $\hat{\beta}^* = a\hat{\beta} + (1-a)\tilde{\beta}$, with a given by

$$a = \frac{\tilde{\tau}_{22} - \tilde{\tau}_{12}}{\hat{\tau}_{11} - 2\tilde{\tau}_{12} + \tilde{\tau}_{22}},$$

since it yields the convex combination with minimal asymptotic variance. The sampling variance of $\hat{\beta}^*$ is estimated by

$$\widehat{\text{var}}\hat{\beta} = \frac{\hat{\tau}_{12}\tilde{\tau}_{22} - \tilde{\tau}_{12}^2}{\hat{\tau}_{11} - 2\tilde{\tau}_{12} + \tilde{\tau}_{22}},$$

on which a confidence interval can be based in the usual way.

3. The Main Theorem

We now state and prove the main theorem, which essentially supports the usual truncate-by-eye approach.

Theorem 3.1. *Let $\beta > 1$ be fixed and $X_1, X_2, \ldots \overset{\text{iid}}{\sim} f(x) = cx^{-\beta}(x = 1, 2, \ldots)$. Define $H_n(x) = \sum_{i=1}^{n} 1\left[X_i = x\right]$ and $M_n = \min\{x : H_n(x) = 0\}$. Provided a sequence y_n satisfies the growth condition*

$$y_n e^{-ncy_n^{-\beta}} \longrightarrow 0 \tag{3.1}$$

as $n \to \infty$, it follows that $\Pr\left[M_n > y_n\right] \to 1$ as $n \to \infty$.

Remark 3.2. (a) The conclusion of the theorem can be expressed by calling y_n a "sequence of probable lower bounds for M_n."

(b) The growth condition on y_n is implicit and awkward for direct checking. Explicit sufficient conditions are developed after the proof. They generally entail $y_n \to \infty$, sublinearly, but still enclosing a fraction of the data $\{X_i\}$ tending to 1.

Proof. First we "poissonize" the sample size n: For each n, let

$$N = N_n \sim \text{Poisson}(n),$$

independently of $\{X_i\}$, and let $M_n^* = M_N$. We prove that $\Pr[M_n^* > y_n] \to 1$ when y_n satisfies the growth condition (3.1), and then "depoissonize" at the end.

Fix a positive integer y. Then

$$\begin{aligned}\Pr\left[M_n^* > y\right] &= \prod_{x=1}^{y} \Pr\left[H_N(x) > 0\right]\\ &= \prod_{x=1}^{y} \mathrm{E}\left[\Pr\left(H_N(x) > 0 | N\right)\right]\\ &= \prod_{x=1}^{y} \mathrm{E}\left[1 - (1 - f(x))^N\right]\\ &= \prod_{x=1}^{y} 1 - e^{-nf(x)}.\end{aligned}$$

The last equality follows from the moment-generating function of the Poisson random variable; i.e., for any $t \in \mathbb{R}$,

$$\mathrm{E}\left[e^{Nt}\right] = e^{n(e^t - 1)}.$$

Now fix an integer sequence $y_n \to \infty$. To have $\Pr\left[M_n^* > y\right] \to 1$, it is enough to argue that

$$\sum_{x=1}^{y_n} \log\left(1 - e^{-nf(x)}\right) \longrightarrow 0.$$

Extend the integer argument of $f(x) = cx^{-\beta}$ ($c = \zeta(\beta)^{-1}$) to a real argument x. Then we have the following bound:

$$\begin{aligned}\left|\sum_{x=1}^{y_n} \log\left(1 - e^{-nf(x)}\right)\right| &< 2\sum_{x=1}^{y_n} e^{-ncx^{-\beta}}\\ &< 2\int_1^{y_n+1} e^{-ncx^{-\beta}}\, dx\\ &< 2y_n e^{-nc(y_n+1)^{-\beta}},\end{aligned}$$

where the last bound tends to zero if the growth condition (3.1) of the theorem is satisfied; this is seen by the following inequality chain:

$$y_n e^{-ncy_n^{-\beta}} < y_n e^{-nc(y_n+1)^{-\beta}} < (y_n+1)e^{-nc(y_n+1)^{-\beta}}.$$

The following depoissonization argument then completes the proof. Let $0 < \delta < 1/2$ and let integers $y_n \to \infty$ satisfy the growth condition (3.1). Then

$$\begin{aligned} P\left[M_N > y_n\right] &= P\left[M_N > y_n \cap N \le n + n^{1/2+\delta}\right] \\ &\quad + P\left[M_N > y_n \cap N > n + n^{1/2+\delta}\right] \\ &\le P\left[M_{[n+n^{1/2+\delta}]} > y_n\right] + o(1), \end{aligned}$$

where $[\cdot]$ in the subscript of M represents the greatest integer function. Since the left side of the above inequality tends to 1, so does $P\big[M_{[n+n^{1/2+\delta}]} > y_n\big]$. Write

$$n + n^{1/2+\delta} = \nu(n) = n\,(1+o(1)) \quad \text{as } n \to \infty.$$

Inverting, we obtain

$$n(\nu) = \nu^{-1}(\nu) = \nu(1+o(1)) \quad \text{as } \nu \to \infty,$$

so $P\left[M_{[\nu]} > y_{n(\nu)}\right] \to 1$ as $\nu \to \infty$ along values making $n(\nu)$ integers. To finish the argument we need to show that

$$y_n e^{-ncy_n^{-\beta}} \to 0 \quad \text{implies} \quad y_{n(\nu)} e^{-\nu c y_{n(\nu)}^{-\beta}} \to 0 \quad \text{as} \quad \nu \to \infty.$$

Taking a log yields

$$\begin{aligned} \lim_{\nu\to\infty} \log\left(y_{n(\nu)} e^{-\nu c y_{n(\nu)}^{-\beta}}\right) &= \lim_{\nu\to\infty}\left[\log y_{n(\nu)} - n(\nu)\,(1+o(1))\,c y_{n(\nu)}^{-\beta}\right] \\ &= \lim_{n\to\infty}\left[\log y_n - n(1+o(1))c y_n^{-\beta}\right] \\ &= -\infty, \end{aligned}$$

as required, since $\log y_n \to +\infty$ and the other term dominates. □

4. An Explicit Growth Condition Sufficient for the Main Theorem

The first and crudest argument below leads to an explicit growth condition that works for all β. We then improve it in the range $\beta \ge 2$ and then extend the improvement to $\beta \ge 1.84$ (which could be further improved with difficulty and perhaps little practical gain).

Fix $\gamma > 1/\beta$; the closer to equality, the weaker the condition and the faster the guaranteed growth of y_n (although the price may be that asymptotics are slow to take hold). We show that

$$y_n = o\left(n^{1/\beta}(\log n)^{-\gamma}\right)$$

makes $y_n e^{-ncy_n^{-\beta}} \to 0$, as the theorem requires. The argument rests on showing that $ncy_n^{-\beta} - \log y_n \to \infty$. With our choice of y_n, the left side exceeds

$$ncy_n^{-\beta} - \frac{1}{\beta}\log n + \gamma \log\log n > c\,(\log n)^{\beta\gamma} - \frac{1}{\beta}\log n + \gamma\log\log n \longrightarrow \infty,$$

since $\beta\gamma > 1$. If $\beta \geq 2$, a further improvement is possible: simply take any

$$y_n = o\left(\left(\frac{n}{\log n}\right)^{1/\beta}\right).$$

That is, let γ above equal $1/\beta$, and the bound grows faster. This follows from examining again

$$ncy_n^{-\beta} - \log y_n > \left(c - \frac{1}{\beta}\right)\log n + \frac{1}{\beta}\log\log n,$$

which tends as it must to infinity, provided $c \geq 1/\beta$. But

$$c = \frac{1}{\sum_{x=1}^{\infty} x^{-\beta}} > \frac{1}{1 + \int_1^{\infty} \frac{dx}{x^{\beta}}} = 1 - \frac{1}{\beta},$$

showing that $c \geq 1/\beta$ as long as $\beta \geq 2$.

For a further improvement, observe that the 1 in the denominator can be replaced by

$$\zeta(2) - 1 = \frac{\pi^2}{6} - 1 = \sup_{s\in(1,2]}\left[\zeta(s) - \int_1^{\infty} x^{-s}\,dx\right]. \tag{4.1}$$

(The last assertion, involving the supremum of the zeta function with its pole $\frac{1}{s-1} = \int_1^{\infty} x^{-s}\,dx$ removed, is nontrivial and is argued in Section 5.) Finally, after solving a quadratic, a permissible range for β is found to be

$$\beta \geq \frac{1}{12}\left(\pi^2 + \sqrt{288 - 24\pi + \pi^4}\right) \approx 1.838,$$

improving (but not optimally) on $\beta \geq 2$.

In summary, explicit sufficient growth conditions for y_n are given by

$$\begin{aligned} 1 < \beta < 1.84: &\quad y_n = o\left(\left(\frac{n}{(\log n)^{1+\delta}}\right)^{1/\beta}\right) \quad (\delta > 0 \text{ fixed, arbitrary}),\\ \beta \geq 1.84: &\quad y_n = o\left(\left(\frac{n}{\log n}\right)^{1/\beta}\right). \end{aligned}$$

The conditions are "close" to necessary as well, in that the growth condition of the theorem implies by an easy argument $y_n = O\left(n^{1/\beta}\right)$.

5. A Proof of the Growth Condition in the Main Theorem

We proceed to show that

$$g(s) = \zeta(s) - \int_1^\infty x^{-s}\,dx = \zeta(s) - \frac{1}{s-1}$$

is strictly increasing on $(1, 2]$, which will justify the assertion

$$\zeta(2) - 1 = \sup_{s \in (1,2]} \left[\zeta(s) - \int_1^\infty x^{-s}\,dx\right].$$

The Riemann zeta function has the following Laurent series expansion about $s = 1$ [Havil 09, p. 118]:

$$\zeta(s) = \frac{1}{s-1} + \sum_{n=0}^\infty \frac{(-1)^n}{n!}\gamma_n (s-1)^n,$$

where the constants γ_n are referred to as the Stieltjes constants. We have the following bounds on γ_n [Berndt 72]:

$$|\gamma_n| < \begin{cases} \frac{4(n-1)!}{\pi^n}, & n \text{ even},\\ \frac{2(n-1)!}{\pi^n}, & n \text{ odd}. \end{cases}$$

We wish to show that for $s \in (1, 2]$,

$$g'(s) = \sum_{n=1}^\infty \frac{(-1)^n}{n!}\gamma_n n (s-1)^{n-1} > 0.$$

Note that

$$\begin{aligned} g'(s) &= -\gamma_1 + \gamma_2(s-1) - \frac{\gamma_3}{2}(s-1)^2 + \sum_{n=4}^{\infty} \frac{(-1)^n}{n!}\gamma_n n(s-1)^{n-1} \\ &> -\gamma_1 + \gamma_2(s-1) + -\frac{\gamma_3}{2}(s-1)^2 - \sum_{n=4}^{\infty} \frac{4(n-1)!}{\pi^n n!} n(s-1)^{n-1} \\ &= -\gamma_1 + \gamma_2(s-1) + -\frac{\gamma_3}{2}(s-1)^2 - \frac{4}{\pi}\sum_{n=4}^{\infty} \left(\frac{s-1}{\pi}\right)^{n-1} \\ &= -\gamma_1 + \gamma_2(s-1) + -\frac{\gamma_3}{2}(s-1)^2 - \frac{4(s-1)^3}{\pi(1+\pi-s)}. \end{aligned}$$

Using the bounds

$$\begin{aligned} \gamma_1 &= -0.07281584548\ldots < -0.072, \\ \gamma_2 &= -0.009690363192\ldots > -0.01, \\ \gamma_3 &= 0.002053834420\ldots < 0.003, \end{aligned}$$

and

$$\left.\frac{4(s-1)^3}{\pi(1+\pi-s)}\right|_{s=2} = \frac{4}{\pi^3(\pi-1)} = 0.0602384\ldots < 0.0603,$$

we see that for all $s \in (1, 2]$,

$$g'(s) > 0.072 - 0.01 - \frac{0.003}{2} - 0.0603 = 0.0002 > 0.$$

This bound can certainly be improved by utilizing the stronger bound of the Stieltjes constants on odd indices and also by isolating more terms before applying the Berndt bound.

References

[Barabási and Albert 99] A. L. Barabási and R. Albert. "Emergence of Scaling in Random Networks." *Science* 286:5439 (1999), 509.

[Barabási and Bonabeau 03] A. L. Barabási and E. Bonabeau. "Scale-Free Networks." *Scientific American* 288:5 (2003) 50–59.

[Berndt 72] B. C. Berndt. "On the Hurwitz Zeta-Function." *Rocky Mountain J. Math.* 2:1 (1972), 151–157.

[Havil 09] J. Havil. *Gamma: Exploring Euler's Constant.* Princeton: Princeton University Press, 2009.

[Newman 01] M. E. J. Newman. “The Structure of Scientific Collaboration Networks.” *Proceedings of the National Academy of Sciences* 98:2 (2001), 404.

[Ravasz and Barabási 03] E. Ravasz and A. L. Barabási. “Hierarchical Organization in Complex Networks.” *Physical Review E* 67:2 (2003), 26112.

[Watts and Strogatz 98] D. J. Watts and S. H. Strogatz. “Collective Dynamics of Small-World Networks.” *Nature* 393 (6684) (1998), 440–442.

Ian Abramson, Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 (abramson@ucsd.edu)

Arthur Berg, Division of Biostatistics, Pennsylvania State University, Hershey, PA 17033 (berg@psu.edu)

Received September 23, 2009; accepted April 14, 2010.