9 Power laws

9.1 Introduction to power laws

The mathematical models of biological systems I considered so far are *conceptual* in some sense. That is, they describe qualitative phenomena, and less suitable for quantitative description of available data. This does not mean, of course, that the models I discussed can be used for a quantitative analysis (e.g., it is a fact that the exponential growth does describe the population increase when the supply is virtually unlimited, and the logistic equation can be used to estimate the population carrying capacity in some cases), but such applications should be performed with great care and understanding of a huge amount of simplifying assumptions put into these models. Sometimes, however, our models should be able to describe an observed phenomenon not only at a qualitative level, but also quantitatively. In this lecture I plan to discuss one such phenomenon and possible mathematical models explaining it. Let me start with a biological example.

In biological classification *species* is the lowest possible rank, and species are grouped together in *genera* (or, singular, *genus*). The next taxonomic unit is a *family*. Therefore, we can talk about the distribution of the number of species in different genera in a given family. Such data can be collected and analyzed. An example of such data is given in the figure below (this example is borrowed from a wonderful paper by G. Yule[1])

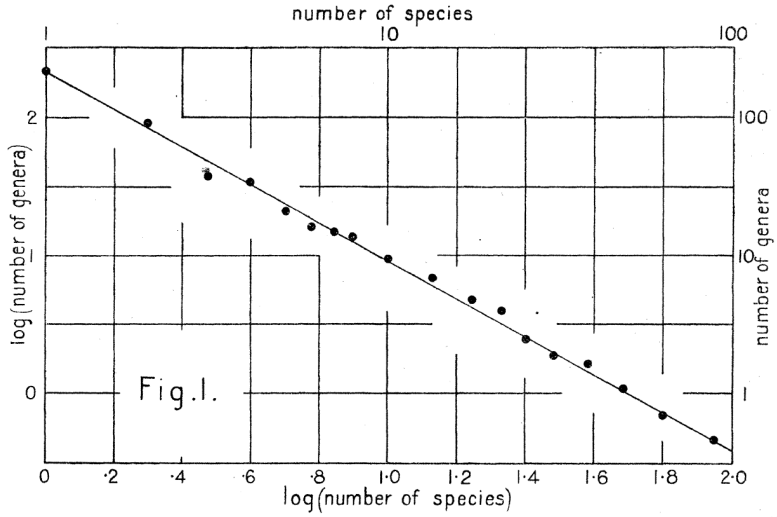

Fig. 1.—Double logarithmic chart for the frequency distribution of sizes of genera in the *Chrysomelidæ*: logarithm of the number of genera plotted on the vertical to logarithm of the number of species on the horizontal. Data in Appendix, Table A.

Here a *distribution* of the sizes of the genera is shown for the family *Chrysomelidæ*. The distribution means that the number of genera with one species was counted, with two species, with three species, and so on, and after it these numbers were plotted against species numbers (to be more precise, for genera with more than 9 species actually intervals of number of species were considered, i.e., the

Math 484/684: Mathematical modeling of biological processes by Artem Novozhilov
e-mail: artem.novozhilov@ndsu.edu. Spring 2014

[1] Yule, G.U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character, 213(402-410), 21-87.

number of genera with 9–11 species, with 12–14 species, with 15–20 species, and so on, the data can be found in the cited paper). Also note that both axes in the figure have logarithmic scaling, i.e., the natural logarithms of the number of genera are plotted versus the natural logarithms of the number of species in a given genera. The observations follows a very simple pattern, namely, in double logarithmic coordinates they are very close to a strait line. That is, if I denote $p(x)$ the number of genera (or sometimes it is more convenient to talk about frequency, i.e., the number divided by the total number of observations) and x the genera size, then I have

$$\log p(x) = a - \alpha \log x,$$

where a and α are positive constants. Therefore, in the usual coordinates

$$p(x) = \frac{A}{x^\alpha}\,. \tag{1}$$

Distributions of the form (1) are said to follow a *power law*. It is a very surprising fact (well, for me at least) how many quantities around us follow a power law asymptotically, starting from some x_0. The constant α is called the *exponent* of the power law.

Here is another example. Consider a distribution of the cites in USA with population more than 40000. There are slightly over 1000 such cities, and it is clear that the number of cities with large populations is very small (we cannot have more than 40 cities of the size of New York in USA). You can see a histogram of this distribution in the figure, By histogram I mean that I took the whole interval from 4×10^5 to 8.2×10^6 (the population of New York), divided into 50 intervals or bins (50 is taken as an example, a good rule of thumb is to take the number of intervals as $\sqrt{N}$, where N is the number of observations, in my case $N = 1036$) of equal length, and counted the number of cities in each interval, after that I plot these numbers versus the centers of the intervals. You can see that both attempts to present the data are not very successful: In the first graph the number of cities with relatively small populations dominates the figure, in the second figure the fact that the number of cities with large populations is very small brings a lot of noise.

To overcome this difficulty we can follow G.Yule and make logarithmic binning, i.e., I divide not into the intervals of equal length, but such that the length of the second interval β times bigger

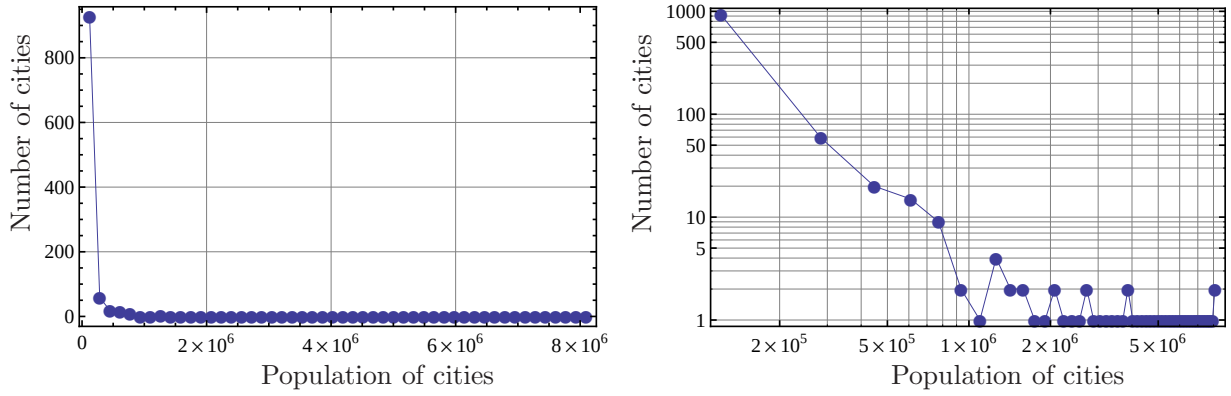

Figure 1: The distribution of the city population in USA for the cities with more than 40000 people. Left: Usual coordinates, right: double logarithmic coordinates

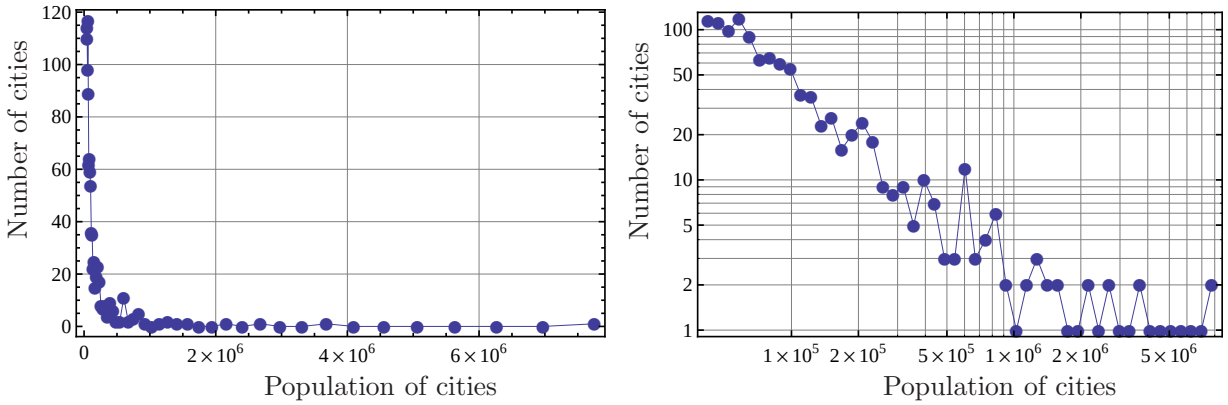

Figure 2: The distribution of the city population in USA for the cities with more than 40000 people. Left: Usual coordinates, right: double logarithmic coordinates. In both cases the logarithmic binning is used, number of bins 50

than the first one, where β is a suitably chosen constant, and so on. You can the result of these manipulations in the figure. This time the result is better, especially in the logarithmic coordinates, but still the histograms contain a lot of obvious noise, especially for large cities.

Let me change the number of bins (intervals) from 50 to 11. Here what I get:

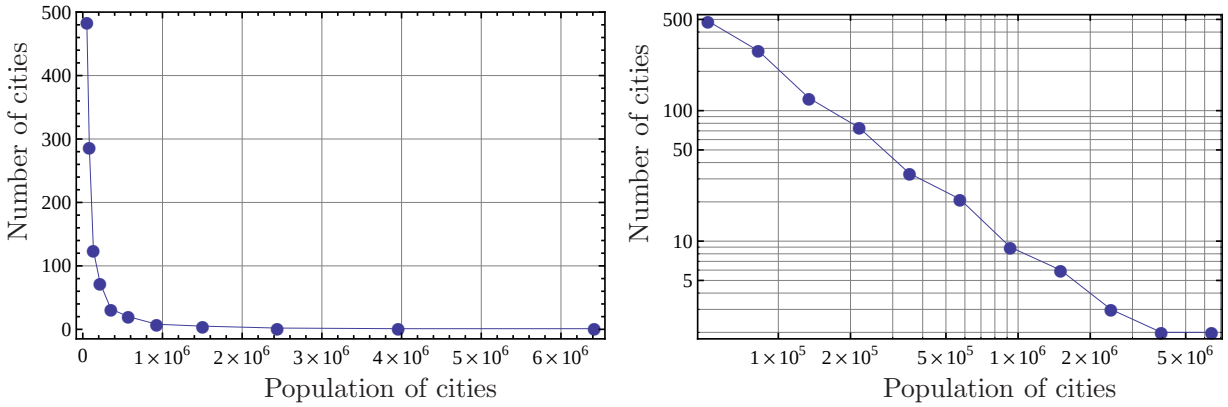

Figure 3: The distribution of the city population in USA for the cities with more than 40000 people. Left: Usual coordinates, right: double logarithmic coordinates. In both cases the logarithmic binning is used, number of bins is 11

Therefore, it is quite reasonable to formulate a hypothesis that the city population distribution follows a power law.

Instead of playing with the number of bins and different binning strategies, it is often more convenient to represent the data in a different way. Namely, consider again $p(x)$, which I define now as the proportion of the quantity of interest in the interval from x to $x + \mathrm{d}x$, where $\mathrm{d}x$ is sufficiently small.

Since $p(x)$ gives a proportion, I should have

$$\int_{x_{\min}}^{\infty} p(x)\,\mathrm{d}x = 1,$$

where $x_{\min} > 0$, because otherwise the integral will diverge. In the language of probability theory, my $p(x)$ is called the *probability density function*. The chance that a given randomly picked city will have a population from x_1 to x_2 (note that it does not make much sense to talk that a city has a population exactly 45672, we need to provide an interval to get a meaningful answer) is given then

$$\mathrm{P}\left\{x_1 \le x \le x_2\right\} = \int_{x_1}^{x_3} p(x)\,\mathrm{d}x,$$

and here I use the notation $\mathrm{P}\left\{A\right\}$ to denote the probability (chance) of event A. Now instead of $p(x)$ consider function

$$F(x) = \mathrm{P}\left\{\text{Population of a city not less than } x\right\} = \int_{x}^{\infty} p(\xi)\,\mathrm{d}\xi = \frac{A}{\alpha-1}x^{-(\alpha-1)} = \left(\frac{x}{x_{\min}}\right)^{-\alpha+1},$$

which also follows the power law (here I found A such that $\int_{x_{\min}}^{\infty} p(x)\,\mathrm{d}x = 1$). The advantage of $F(x)$ is that this function is well defined for any x given the data, we simply need to calculate the proportion (or the number) of cities, whose population is bigger than x. If $F(x)$ follows the power law, therefore the data should have the power law distribution. The graph of $F(x)$ is often called *rank* or *frequency plot*, because $F(x)$ is proportional to the rank of x (see the figure). Here we see exactly the same pattern that the data in double logarithmic coordinates can be describes by a straight line, and hence the data themselves follows the power law.

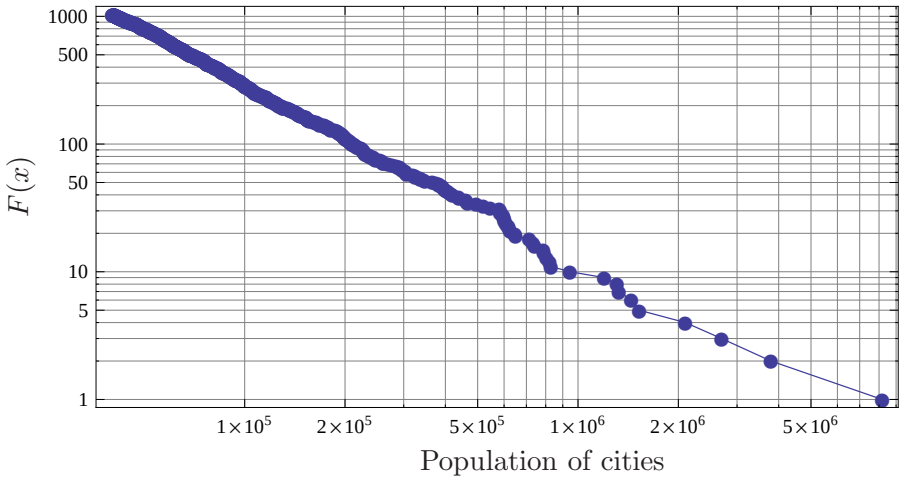

Figure 4: The function $F(x) = \#\{\text{Cities with population not less than } x\}$ versus the city population in double logarithmic scale

To impress you even more, consider now the word count is a book (I picked *On the Origin of Species* by Charles Darwin, but a similar picture can be observed for other texts), and consider the distribution of the counts in a text, i.e., how many words were used only once, how many words were

used twice, three times, and so on. It is quite clear that there should be a significant number of words that were used only few times, and not so many words that are used in almost every paragraph, but the actual distribution of these numbers?

Just for curious, here are 10 words that appear most often:

the, of, and, in, to, a, that, have, be, as

and here are several examples of the words that appear only once:

last, decision, personal, drew, patiently, 1837, philosophers, mysteries, INTRODUCTION.

The function $F(x)$, which gives the number of words used not less than x times can be plotted (see the figure), and the result, which shows that is not unreasonable to put forward the hypothesis of the power law, is known in linguistics as *Zipf's law*.

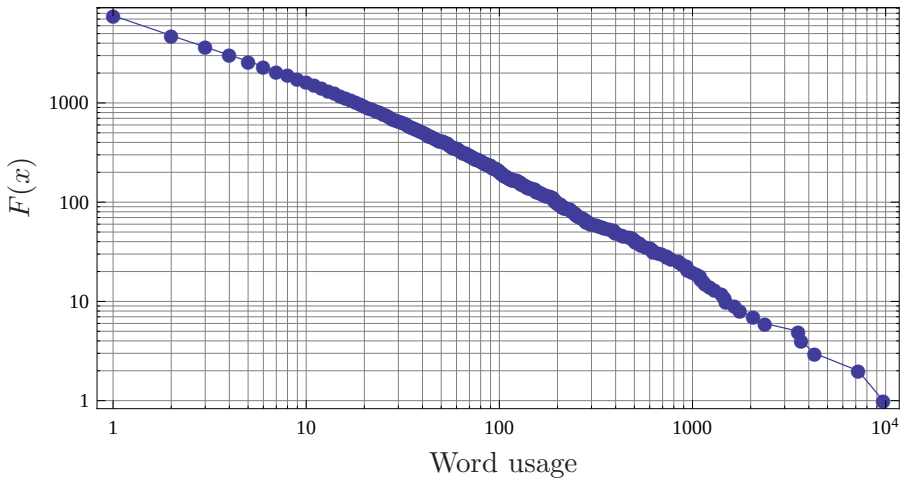

Figure 5: The function $F(x) = \#\{\text{Words used more than } x \text{ times}\}$ versus x in double logarithmic scale

The list of examples can be easily continued. In particular, the available data suggest that the following quantities obey the power law: citations of scientific papers, copies of books sold, magnitude of earthquakes, intensities of wars, wealth of reaches Americans, frequencies of family names, etc (for those interested to read much more about power laws I recommend this paper[2]).

Here are a few points about power laws. In the literature a more general definition of the power law distribution is used. A quantity X is said to have a power law distribution with exponent α if it has a probability density function

$$p(x) \sim Ax^{-\alpha},$$

i.e., if $p(x)$ behaves similarly to the exact power law asymptotically, for large x. Since one assumes that usually X takes values in the interval $(x_{\min}, \infty)$, then it is necessary to request that $\alpha > 1$ to guarantee that the integral

$$\int_{x_{\min}}^{\infty} p(x)\,\mathrm{d}x$$

[2]Clauset, A., Shalizi, C.R., & Newman, M.E. (2009). Power law distributions in empirical data. SIAM review, 51(4), 661-703.

converges. Very important quantities that describe X are its moments; two most important are the *expectation* $\mathrm{E}(X)$ (the first moment) and the *variance* $\mathrm{Var}(X)$ (the second central moment), they are defined as

$$\mathrm{E}(X) = \int_{x_{\min}}^{\infty} xp(x)\,\mathrm{d}x, \quad \mathrm{Var}(X) = \int_{x_{\min}}^{\infty} \big(x - \mathrm{E}(X)\big)^2 p(x)\,\mathrm{d}x.$$

The expectation shows the average value of X and the variance shows the spread of X around $\mathrm{E}(X)$, the bigger the variance the less predictable quantity X becomes. A very important fact is that for the variance to exist α has to be bigger than 3 and for the expectation to exist α has to be bigger than 2. It turns out that the estimates of α from available data usually show that $2 < \alpha < 3$, i.e., the observed quantities follow the power law with infinite variance. This is why such distributions usually called *fat-tailed distributions*.

Example 1 (80/20 law)**.** Let me ask the question is where the majority of the distribution of X lies if X has a power law distribution. Consider for example such value m that

$$\int_{m}^{\infty} p(x)\,\mathrm{d}x = \frac{1}{2}\int_{x_{\min}}^{\infty} p(x)\,\mathrm{d}x.$$

m is called the *median* of the distribution. For the power law we can find that

$$m = 2^{1/(\alpha-1)} x_{\min}.$$

For example, if X is the wealth, then m is such income that exactly 50% people have smaller income than m and 50% of people have bigger income than m. But we can also ask how much income lies in these two halves. The fraction of money of the reacher half is given by

$$\frac{\int_{m}^{\infty} xp(x)\,\mathrm{d}x}{\int_{x_{\min}}^{\infty} xp(x)\,\mathrm{d}x} = 2^{-\frac{\alpha-2}{\alpha-1}},$$

provided $\alpha > 2$ and the integrals converge. Therefore, if, say $\alpha = 2.1$ (estimated from data), then the fraction of $2^{-0.091} \approx 0.94$ of the wealth in the hands of the richest half. A generalization can be made that the fraction

$$W(p) = p^{\frac{\alpha-2}{\alpha-1}}$$

of the wealth in the hands of the richest p. Using this expression, we can see that approximately 80% of the wealth in the hands of the richest 20%, hence the famous 80/20 law.

Finishing this section I would like to mention that power law distributions are also very often called *scale-free distributions*. I saw several different explanations in the literature to justify the use of this term, however none of them is very convincing, therefore, it would be a much better practice to stick to more informative power law distribution.

9.2 Birth–death–innovation model of protein domain evolution

Proteins consists of *domains*, that can be defined as structural elements of a given protein that can function and evolve independently of the rest of the protein chain. One of the main mechanisms of genome evolution is *duplication*, which corresponds to appearance of two copies of the same sequence of DNA in genome. If a particular DNA sequence corresponds to a protein domain, then after duplication

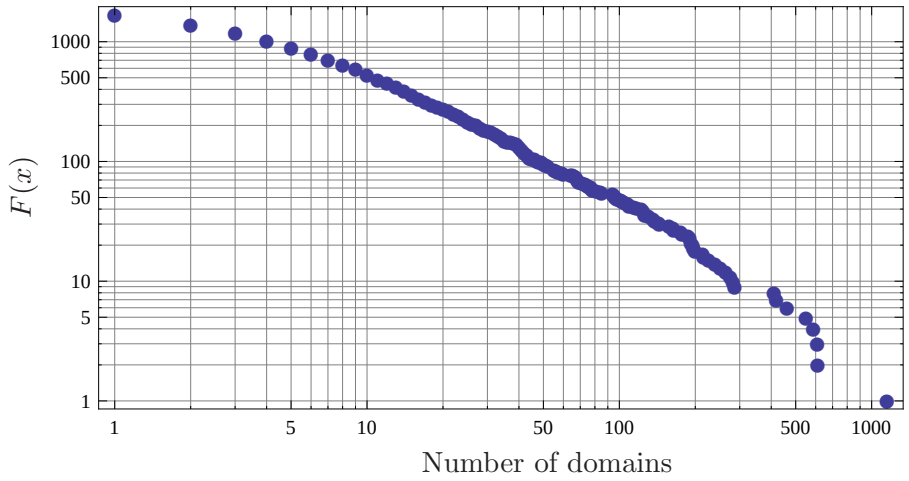

Figure 6: The function $F(x) = \#\{\text{Families with not less than } x \text{ domains}\}$ versus x in double logarithmic scale for *Homo Sapience*

of this sequence there will be two similar domains that belong to the same family. It turns out that the distribution of domain families follows asymptotically the power law (see the figure).

In this section I present a mathematical model of the domain evolution that produces power law with exponent $\alpha = 1$.

Consider possible elementary events in genome evolution:

- *Domain birth.* This is usually caused by duplication and forces a family with k domains becomes the family with $k + 1$ domains;
- *Domain death.* This is usually caused by inactivating mutations and results in a family with $k - 1$ domains if there were originally k domains;
- *Domain innovation.* This event gives birth to a new domain family with 1 domain. This can happen for several reasons, one of which mutation in one of the existing domain such that the domain stays functional, but now it is so different from the rest of his siblings that we count it as a new domain family.

Assume that it is possible to have maximum d domains in a family, and let x_k be the number of families with k domains, $k = 1, \ldots, d$. Assume that the rates of duplications and inactivating mutations are constant per one domain and denote them λ and μ respectively. We have that the rate of change of the number of domain families with only one domain is given by

$$\dot{x}_1 = -(\lambda + \mu)x_1 + 2\mu x_2 + \nu.$$

Here $-\lambda x_1$ term corresponds to duplication that makes one of such families a family with two domains, $-\mu x_1$ term corresponds to possible mutation that destroys one family, $2\mu x_2$ describes possible inactivating mutations in the families with two domains, and coefficient 2 comes from the fact that each family among x_2 has 2 domains, finally μ denotes a constant rate of domain innovations. For the

rest of the variables we have

$$\dot{x}_k = (k-1)\lambda x_{k-1} - k(\lambda+\mu)x_k + (k+1)\mu x_{k+1}, \quad k = 2, \ldots, d-1,$$
$$\dot{x}_d = (d-1)\lambda x_{d-1} - d\mu x_d.$$

It is convenient to write the system of ODE in the vector form

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\nu}, \tag{2}$$

where the matrix $\boldsymbol{A}$ is given by

$$\begin{bmatrix} -(\lambda+\mu) & 2\mu & 0 & \ldots & 0 \\ \lambda & -2(\lambda+\mu) & 3\mu & \ldots & 0 \\ 0 & 2\lambda & \ddots & \ldots & \ldots \\ \vdots & \vdots & \vdots & -(d-1)(\lambda+\mu) & d\mu \\ 0 & 0 & \ldots & (d-1)\lambda & -d\mu \end{bmatrix},$$

and the vector $\boldsymbol{\nu} = (\nu, 0, \ldots, 0)^\top$. One can check that $\det \boldsymbol{A} = (-1)^d k! \mu^d \neq 0$ if $\mu \neq 0$, and hence there is unique equilibrium $\hat{\boldsymbol{x}}$, which is the solution to

$$\boldsymbol{A}\boldsymbol{x} = -\boldsymbol{\nu}.$$

The stability of this equilibrium is determined by the real parts of the eigenvalues of $\boldsymbol{A}$ because the change of variables $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{A}^{-1}\boldsymbol{\nu}$ leads to the system that we studied: $\dot{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}$.

Theorem 2. *Consider system* (2) *and let* $\theta = \lambda/\mu$. *Then equilibrium* $\hat{\boldsymbol{x}}$ *is asymptotically stable and given by*

$$x_k = \frac{\nu}{\lambda}\frac{\theta^k}{k}, \quad k = 1, \ldots, d.$$

and has the power law form with exponent 1 if and only if $\theta = 1$, *i.e., if* $\lambda = \mu$.

To prove this theorem we actually need first to establish that the coordinates of the equilibrium exactly as given and that matrix $\boldsymbol{A}$ has eigenvalues with negative real parts. For the second part I will need the following fact (without proof):

Proposition 3. *Consider matrix* $\boldsymbol{A}$ *and its principal minors* M_k, $k = 1, \ldots, d$ *(i.e., the determinants of the matrices obtained from* $\boldsymbol{A}$ *by keeping first* k *columns and rows). Matrix* $\boldsymbol{A}$ *is asymptotically stable (i.e., for all eigenvalues it is true that* $\operatorname{Re}\lambda < 0$*) if and only if* $(-1)^k \boldsymbol{M}_k > 0$ *for all* $k = 1, \ldots, d$.

Proof of Theorem 2. First, by summing all the equations in (2), we find that

$$-\mu x_1 + \nu = 0 \implies \hat{x}_1 = \frac{\nu}{\mu}.$$

Now from the second equation in (2) we can find $\hat{x}_2$, from the third $\hat{x}_3$ and so on, which proves the formulae for the coordinates of the equilibrium.

We have that $M_1 = -(\lambda+\mu)$, and $M_d = \det \boldsymbol{A} = (-1)^d d! \mu^d$ (the last expression can be found by first adding the last row of $\boldsymbol{A}$ to row $d-1$, after that row $d-1$ added to $d-2$ and so on, eventually we will get a diagonal matrix with $(-\mu, -2\mu, \ldots, -d\mu)$ on the main diagonal).

Now consider matrices $\boldsymbol{M}_k$ such that $M_k = \det \boldsymbol{M}_k$, $k = 1, \dots d-1$. By using the determinant formulae through the last row, I find

$$M_k = \det \boldsymbol{A}_k - k\lambda M_{k-1},$$

where $\boldsymbol{A}_k$ is $k \times k$ matrix of the same structure as $\boldsymbol{A}$ and whose $\det \boldsymbol{A}_k = (-1)^k k! \mu^k$. We know M_1, and hence can find M_2 and so on. The general formulae is given by

$$M_k = (-1)^k k! \frac{\lambda^{k+1} - \mu^{k+1}}{\lambda - \mu},$$

which proves that all the eigenvalues of $\boldsymbol{A}$ have negative real parts. ■

The biggest question of course is how to modify the model to obtain a power law distribution with exponent other than one. For this one can assume that the birth and death rates are not constant by depend on the size of the given family. E.g., we can assume that the birth rate is λ_k and the death rate is μ_k per domain family of the size k (we retrieve the model, which was analyzed above, if we set $\lambda_k = k\lambda$, $\mu_k = k\mu$). The following theorem holds.

Theorem 4. *Birth–death–innovation model with the size dependent birth and death rates has a unique asymptotically stable equilibrium $\hat{\boldsymbol{x}} \in \mathbf{R}^d$ with coordinates*

$$\hat{x}_i = \frac{\nu \prod_{k=1}^{i-1} \lambda_k}{\prod_{k=1}^{i} \mu_k}, \quad i = 1, \dots, d.$$

This equilibrium distribution follows asymptotically the power law distribution with exponent α if and only if

$$\frac{\lambda_{k-1}}{\mu_k} = 1 + \frac{\alpha}{k} + \mathcal{O}(k^{-2}).$$

For example, according to this theorem, the model with $\lambda_k = \lambda(k+a)$, $\mu_k = \mu(k+b)$ for some constants $a \neq b$ will produce a power law with the exponent $a - b - 1$ if $\lambda = \mu$.

9.3 Preferential attachment

Here I present a heuristic derivation of the power law distribution via the so-called principle of *preferential attachment* (in this and the following subsections I follow mainly this paper[3]). Consider the World Wide Web, with can be represented as a directed graph, where the vertices correspond to the web pages and there is an edge from vertex i to vertex j if there is a hyperlink from page i to page j. Now each page can be characterized by the number of hyperlinks to this page, in the graph theoretic language this number is called the *in-degree*. Therefore, we can talk about the in-degree distribution and it turns out that this distribution can be closely approximated by a power law. Assume that a new web page is created. It is reasonable to expect that this new page will link to some popular web pages, i.e., the chance that a new web page is connected to a web-page with in-degree k should be proportional to k, and this is what is usually called the preferential attachment principle.

[3]Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. Internet mathematics, 1(2), 226-251.

Here is an informal argument to formalize the preferential attachment. Let $x_j(t)$ be the number of web pages with in-degree j when there are t pages total. Then, for $j \geq 1$ the probability that $x_j(t)$ increases is simply

$$\alpha \frac{x_{j-1}(t)}{t} + (1-\alpha)\frac{(j-1)x_{j-1}(t)}{t}\,,$$

if we assume that new web page appears with only one link to existing pages, and this one link is chosen randomly among all t pages with probability α and with probability $1-\alpha$ this one link is chosen randomly but with probabilities proportional to the existing in-degrees. Similarly, the probability that $x_j(t)$ decreases is

$$\alpha \frac{x_j(t)}{t} + (1-\alpha)\frac{jx_j(t)}{t}\,.$$

Therefore, for $j \geq 1$,

$$\dot{x}_j = \frac{\alpha(x_{j-1} - x_j) + (1-\alpha)\big((j-1)x_{j-1} - jx_j\big)}{t}\,.$$

The case $j = 0$ should be treated differently since each new web page has in-degree 0, and therefore

$$\dot{x}_0 = 1 - \frac{\alpha x_0}{t}\,.$$

We obtained a non-autonomous system of linear ordinary differential equations. Since time unit in the model is appearance of one new web page, we can assume that the limiting stationary state should have the form

$$x_j(t) = c_j t,$$

where c_j is a constant, which specifies which fraction of the total number of pages the pages with in-degree j constitute.

We have for x_0

$$\dot{x}_0 = c_0 = 1 - \alpha c_0 \implies c_0 = \frac{1}{1+\alpha}\,.$$

For general j

$$c_j(1 + \alpha + j(1-\alpha)) = c_{j-1}(\alpha + (j-1)(1-\alpha)).$$

We can determine c_j exactly using the above recurrence, but for our goal it is enough to note that

$$\frac{c_j}{c_{j-1}} = 1 - \frac{2-\alpha}{1+\alpha+j(1-\alpha)} \sim 1 - \left(\frac{2-\alpha}{1-\alpha}\right)\frac{1}{j}\,.$$

This yields that asymptotically

$$c_j \sim C j^{-\frac{2-\alpha}{1-\alpha}}\,,$$

for some constant C. To see this, note that the last expression means

$$\frac{c_j}{c_{j-1}} \sim \left(\frac{j-1}{j}\right)^{\frac{2-\alpha}{1-\alpha}} \sim 1 - \left(\frac{2-\alpha}{1-\alpha}\right)\frac{1}{j}\,,$$

as required.