Analysis of Multi-Disciplinary Problems

Modelling Scale-Free Properties in Free Software: A Study on *Debian*

This is a set of guidelines for you to follow while you are working on the second numerical assignment. You may recall that in the first assignment you had to work with data pertaining to the growth of *IBM*. The trend that you found in that case was one of an exponential growth, saturated by nonlinearity, modelled readily with the help of the logistic equation. Now in this second assignment, you will work with data taken from the repository of the free-software Operating System, *Debian*. Once again the purpose behind this exercise is to give you a hands-on experience in modelling real data, but as opposed to working with the *IBM* data, in this case, the mathematical principle is different. While the modelling in the case of *IBM* was based on an autonomous equation, in this instance it will be a non-autonomus one. Recalling the discussion regarding scale-free networks should be helpful here. The following items give a description of the contents of the other files in the folder, along with some suggestions for the mathematical modelling.

- **The data files:** There are two of them, named `in_squeeze.dat` and `out_squeeze.dat`. You need to understand the history of *Debian* to follow the contents of these two data files. Since its inception, *Debian* has been undergoing periodic releases of upgraded software. These releases are designed by a team of free-software enthusiasts. Release number 6 has been named *Squeeze*, and the two data files here pertain to the *Squeeze* release.

  To understand what these two data files contain, you will have to view the entire *Debian* software repository as a network of various functional packages. The networking is due to the interdependencies of functions among the packages (the nodes in the network). Now there are two distinct networks among these packages. One is a network of dependency links which are directed towards the packages, while the other is a network of similar links, but directed away from the packages. So the former is an indirected network, while the latter is an outdirected network. The two different data files reflect these two separate properties. The file `in_squeeze.dat` contains data pertaining to the indirected network in the *Squeeze* release of *Debian*, while the file `out_squeeze.dat` contains the corresponding data pertaining to the outdirected network.

  Both files have two columns. The first column gives the number of links, while the second column gives a count of the packages (nodes) having the given number of links. So if you were to plot the first column along the x axis, and the second column along y, what you will effectively have is an unnormalised frequency distribution of the number of packages, for a given number of links.

- **Research articles:** There is a long research article named `nnr_compsys.pdf` (a shorter version of the article is named `nnr_jphys_conf.pdf`) You might read the full document, but of especial importance is the discussion surrounding Equation (2), which you would need for your data-fitting work. Your modelling work on the two data files should result in two separate figures, as shown in Figures 5 & 6. The captions of these two figues will give you a clear idea of the numerical values of the various fixed parameters. The former figure shows the degree distribution of links in the indirected network, while the latter figure shows it for the outdirected network.

- **The mathematical modelling:**

  1. In the research article there is an equation of the form $(x+\lambda)\phi'(x) = \alpha\phi(1-\eta\phi^{\mu})$. The integral solution of this equation is rendered simpler as $\phi(x) = \eta + c^2\,(x+\lambda)^{-2}$, with $\alpha = -2$ and $\mu = -1$. What you have to do is plot the data given in each data file — first column along the x axis, and second column along the y axis — and get a model fit between the data plot and the integral solution. Your guide to the model-fitting exercise should be Figures 5 & 6.
  2. One very important trick that you must adopt is to plot the data in a $\log$-$\log$ format. Only then would you be able to discern the features of a power-law. If there is a straight-line trend anywhere in the $\log$-$\log$ plot, then that portion is governed by a power law. This will be indicated by the value of α, and getting it right in itself should be a triumph. But beyond this, you should also get a good match between the values of μ, η and λ employed by you, and the values indicated in the relevant figure captions.
  3. There are some thoughts for you to take home. What did the value of $\mu = -1$ indicate? You will realise that it can mean one thing only — scale-free properties. Therefore, a power law is very much evident. And you will find that the power-law exponent satisfies Zipf's law, $\alpha = -2$.

     Another interesting thing is that the differential equation leads to an integral solution that is very nearly similar to the standard logistic equation that you are familiar with, and yet, the final results could be so vastly different. This sweeping generalisation is the power of differential equations. You will continue to encounter these simple forms of differential equations again and again in various contexts. The variations will only be due to the values of the specific parameters, and the initial and boundary conditions.