Stat4DS / Homework 01

(Part B)

Pierpaolo Brutti

Due Sunday, November 22, 2020, 23:59 PM on Moodle

General Instructions

I expect you to upload your solutions on Moodle as a **single running** `R Markdown` file (`.rmd`) + its `html` output, named with your surnames.

You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Your responses must be supported by both textual explanations and the code you generate to produce your results. *Just examining your various objects in the "Environment" section of RStudio is insufficient – you must use scripted commands and functions.*

R Markdown Test

To be sure that everything is working fine, start `RStudio` and create an empty project called `HW1`. Now open a new `R Markdown` file (`File > New File > R Markdown...`); set the output to `HTML mode`, press `OK` and then click on `Knit HTML`. This should produce a web page with the knitting procedure executing the default code blocks. You can now start editing this file to produce your homework submission.

Please Notice

- For more info on `R Markdown`, check the support webpage that explains the main steps and ingredients: R Markdown from RStudio.
- For more info on how to write math formulas in LaTex: Wikibooks.
- Remember our **policy on collaboration**: *Collaboration on homework assignments with fellow students is **encouraged**. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had discussions concerning your solution. You may **not**, however, share written work or code after discussing a problem with others. The solutions should be written by **you**.*

Exercise: People have the power (. . . law. . .)

Introduction

In this exercise we focus on the so called **power law** family of distributions. An interesting aspect of these distributions is that, unlike many others we have seen, their variance can be extremely large or even infinite. As a result, certain methods we usually rely on in probabilistic arguments, such as concentration of the sum of random variables (i.e. Chebyshev or Hoeffding inequalities), may **not** apply.

Power laws and related distributions may initially appear surprising or unusual, but in fact they are quite natural, and arise easily in many applied setups. For example, suppose we want to study the *number of times* a word appears in *all* the books printed in English over a year, for example, all across Europe. Some common words, such as "the", "of", and "an", appear remarkably frequently, while most words would only appear at most a handful of times describing an extremely right skewed, very long tailed distribution.

This is just an example. In practice, many other phenomena share this property that the corresponding distribution is not well concentrated around its *mean*, such as the sizes of cities, the strength of earthquakes, the distribution of wealth among families, and the *degree distribution* of real networks (see below). For many such examples, a power law has been shown to provide a very plausible model for their distribution.

Background 1/3: Power Laws

In general, we say that a nonnegative random variable X is said to have a *power law distribution* if

$$\mathbb{P}(X \geqslant x) \sim c \cdot x^{-\alpha}, \quad \text{(complementary CDF)}$$

for constants $c > 0$ and $\alpha > 0$. Here $f(x) \sim g(x)$ means that the limit of the ratio of $f(\cdot)$ and $g(\cdot)$ converges to 1 as $x \to +\infty$.

Example. A Pareto distribution with parameters $\alpha > 0$ and minimum value $m > 0$ – see the `actuar` package – satisfies

$$\mathbb{P}(X \geqslant x) = \left(\frac{x}{m}\right)^{-\alpha} \mathbb{I}_{[m,+\infty)}(x).$$

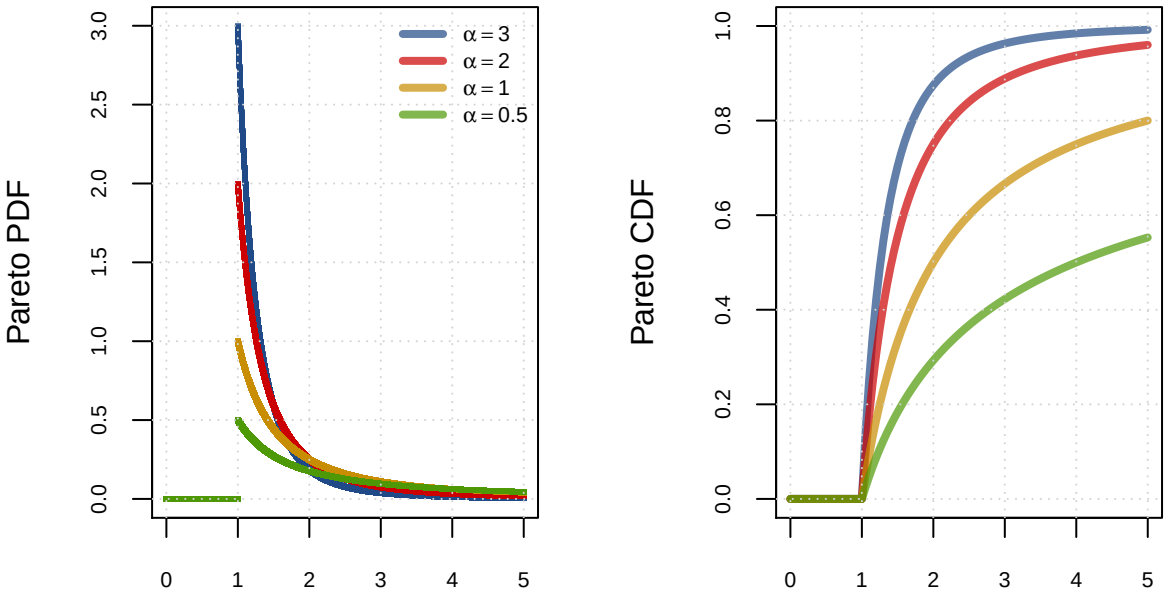
The value α is sometimes called the *tail index*: the *lower* α, the *heavier* the distribution tail is. Its PDF is then,

$$f_X(x) = \alpha\, m^\alpha\, x^{-(\alpha+1)}\, \mathbb{I}_{[m,+\infty)}(x).$$

Let us try to examine the moments of this random variable:

$$\begin{aligned}
\mathbb{E}(X) &= \int_m^{+\infty} x \left(\alpha\, m^\alpha\, x^{-(\alpha+1)}\right) \mathrm{d}x = \alpha\, m^\alpha \int_m^{+\infty} x^{-\alpha} = \begin{cases} +\infty & \text{for } \alpha \leqslant 1, \\ \frac{\alpha\, m}{\alpha - 1} & \text{for } \alpha > 1. \end{cases} \\
\mathbb{E}(X^j) &= \int_m^{+\infty} x^j \left(\alpha\, m^\alpha\, x^{-(\alpha+1)}\right) \mathrm{d}x = \alpha\, m^\alpha \int_m^{+\infty} x^{j-(\alpha+1)} = \begin{cases} +\infty & \text{for } \alpha \leqslant j, \\ \frac{\alpha\, m^j}{\alpha - j} & \text{for } \alpha > j. \end{cases}
\end{aligned}$$

So, for example, when $\alpha \leqslant 2$, the second moment is infinite. Correspondingly, the variance is infinite when $1 < \alpha \leqslant 2$; for $\alpha \leqslant 1$ since both the first and second moments are infinite the variance is not well-defined.



A power law is best visualized on what is called a **log–log plot**, where both axes are presented using logarithmic scales. On a log–log plot the relationship $y = c \cdot x^\alpha$ is shown by presenting $\log(y) = \alpha \cdot \log(x) + \log(c)$, so that the polynomial relationship appears as a straight line whose slope depends on the exponent α. More generally, if X has a power law distribution, then in a log–log plot of the complementary CDF, asymptotically the behavior will be a straight line. It is important to emphasize that the "straight–line" test on a log–log plot is sometimes used to infer that a sample arises from a distribution that follows a power law, but because many other distributions produce nearly linear outcomes on a log–log plot, one must take more care to test for power laws.

Thus far we have focused on the mathematical definitions for *continuous* power law distributions. But we could also consider **discrete** variations. For example, the **zeta distribution** with parameter $s > 1$ is defined for all positive integer values x according to

$$p_X(x) = \mathbb{P}(X = x) = \frac{x^{-s}}{\xi(s)},$$

where the Riemann zeta function $\xi(s)$ is given by $\xi(s) = \sum_{j=1}^{\infty} j^{-s}$. The fact that the PMF $p_X(\cdot)$ is proportional to x^{-s} is the natural discrete analogue for a power law distribution.

Background 2/3: Random Networks Models

Network science aims to build models that reproduce the properties of real networks. Most networks we encounter do **not** have the comforting regularity of a crystal lattice or the predictable radial architecture of a spider web. Rather, at first inspection they look as if they were spun *randomly*. Random network theory embraces this apparent randomness by constructing and characterizing networks that are *truly* random.

From a modeling perspective a network is a relatively simple object, consisting of, say, N **nodes** and (undirected) **links** between them. In our simple setup, a network can be easily represented by an $(N \times N)$ **adjacency matrix** $\mathbb{A}$; that is, by a square, symmetric, binary matrix such that

$$\mathbb{A}[i,j] = \begin{cases} 1 & \text{if there is a link between node } i \text{ and node } j, \\ 0 & \text{otherwise.} \end{cases}$$

The real challenge, however, is to decide where to place the links between the nodes so that we reproduce the complexity of a real system. In this respect the philosophy behind a random network is simple: we assume that this goal is best achieved by placing the links randomly between the nodes. That takes us to two possible definition of a random network:

Definitions (Random Networks).

G(N, L) **Model:** N labeled nodes are connected with L randomly placed links. Erdős and Rényi used this definition in their string of papers on random networks.

G(N, p) **Model:** Each pair of N labeled nodes is connected with probability p, a model introduced by Gilbert.

Hence, the G(N, p) model fixes the probability p that two nodes are connected and the G(N, L) model fixes the total number of links L. In the following we will focus on the G(N, p) model, not only because it's easier to calculate most of its key characteristics, but also because in real networks the number of links rarely stays fixed. Hence, to construct a random network we follow these steps:

Algorithm (Generate a Random Networks).

1. Start with N isolated nodes associated with an *adjacency matrix* $\mathbb{A}$ entirely filled with zeros.
2. Select a node pair and generate a random number between 0 and 1.
3. If this number exceeds p, connect the selected node pair with a link and update $\mathbb{A}$, otherwise leave them disconnected.
4. Repeat step (2) and (3) for each of the $N \cdot (N-1)/2$ node pairs.

In a given realization of a random network some nodes gain numerous links, while others acquire only a few or no links. These differences are captured by the **degree distribution**, which is the probability that a randomly chosen node has degree k.

In a random network the probability that node i has exactly k links is the product of three terms:

- The probability that k of its links are present, or p^k.
- The probability that the remaining $(N-1-k)$ links are missing, or $(1-p)^{N-1-k}$
- The number of ways we can select k links from $(N-1)$ potential links a node can have, or $\binom{N-1}{k}$.

In other words, we are just saying that the random variable K which describes the degree of our random network simply follows a Binom$(N-1, p)$ distribution. From here, we already know that the expected network degree is simply equal to $\mathbb{E}(\mathsf{K}) = p \cdot (N-1)$ with variance $\mathbb{V}\text{ar}(\mathsf{K}) = p \cdot (1-p) \cdot (N-1)$.

But, most real networks are **sparse**, meaning that for them the expected number of link $\mathbb{E}(\mathsf{K}) << N$. Under this "large network" assumption – fixed p, very large N – the degree distribution we just found is well approximated by a **Poisson distribution** with parameter $\mu = \mathbb{E}(\mathsf{K})$:

$$p_k = \mathbb{P}(\mathsf{K} = k) = \frac{\mu^k \, \mathrm{e}^{-\mu}}{k!}.$$

Although interesting, comparison with real data indicates that random network models like this does not capture the degree distribution of real networks. As an example, in random networks, most nodes have comparable degrees, forbidding *hubs* (i.e. nodes with a very large degree), that instead appear to be typical in real networks. To understand why hubs are missing from a random network, we first notice that by Stirling approximation $k! \approx \sqrt{2\pi k}\,\left(\frac{k}{\mathrm{e}}\right)^k$ so that we can rewrite the Poisson PMF above as

$$p_k \approx \frac{\mathrm{e}^{-\mu}}{\sqrt{2\pi k}} \left(\frac{\mathrm{e} \cdot \mu}{k}\right)^k.$$

For degrees $k > \mathrm{e} \cdot \mu$ the term in the parenthesis is smaller than 1, hence for large k both k–dependent terms decrease rapidly with increasing k. Overall the previous equation predicts that in a random network the chance of observing a hub decreases *faster* than exponentially.

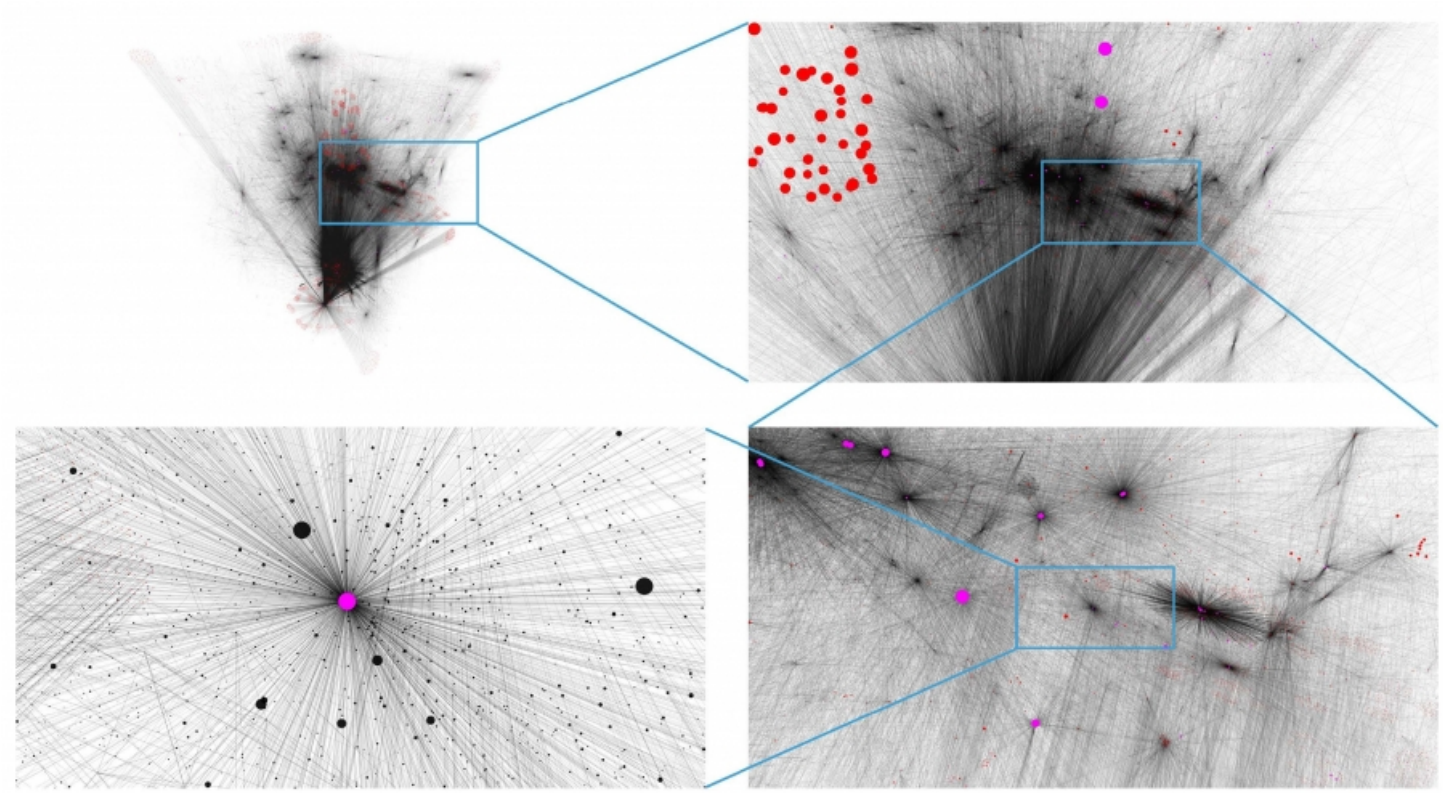
Background 3/3: Real (preferential "attached") Networks

It is difficult to overstate the importance of the World Wide Web in our daily life. Similarly, we cannot exaggerate the role the `WWW` played in the development of network theory: it facilitated the discovery of a number of fundamental network characteristics and became a standard testbed for most network measures.

The `WWW` is a network whose nodes are documents and the links are the *uniform resource locators* (`URLs`) that allow us to "surf" with a click from one web document to the other. With an estimated size of over one trillion documents ($N \approx 10^{12}$), the Web is the largest network humanity has ever built. It exceeds in size even the human brain ($N \approx 10^{11}$ neurons).

The first map of the `WWW` obtained with the explicit goal of understanding the structure of the network behind it was generated by Hawoong Jeong at University of Notre Dame. He mapped out the `nd.edu` domain, consisting of about 300,000 documents and 1.5 million links. The purpose of the map was to compare the properties of the Web graph to the random network model. Indeed, in 1998 there were reasons to believe that the `WWW` could be well approximated by a random network. The content of each document reflects the personal and professional interests of its creator, from individuals to organizations. Given the diversity of these interests, the links on these documents might appear to point to randomly chosen documents.

A quick look at the top–right figure below supports this view: there appears to be considerable randomness. Yet, a closer inspection reveals some striking differences. Indeed, as we said, in a random network *hubs* are effectively forbidden. In contrast below we see that numerous small–degree nodes coexist with a few hubs, nodes with an exceptionally large number of links.

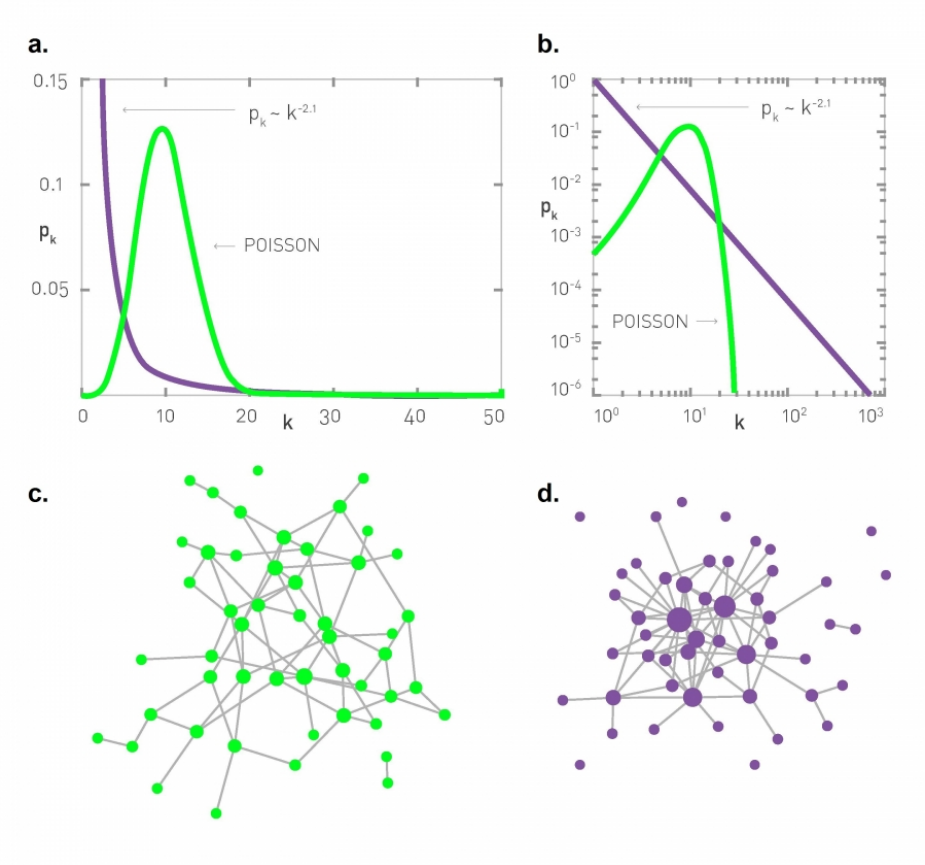


Now, **if** the `WWW` were to be a random net, its degree should follow a Poisson. Yet, as we've seen, the Poisson form offers a very poor fit. Instead it has been observed that on a log–log scale the data points form an approximate straight line, suggesting that the `WWW` degree distribution could be well approximated with one of our beloved **power laws**.

Now the `WWW` is a **directed** network (and its adjacency matrix is not symmetric anymore), hence each document is characterized by an out–degree `kout`, representing the number of links that point from the document to other documents, and an in–degree `kin`, representing the number of other documents that point to the selected document. We must therefore distinguish two degree distributions: the probability that a randomly chosen document points to `kout` web documents, or $p_{\texttt{kout}}$, and the probability that a randomly chosen node has `kin` web documents pointing to it, or $p_{\texttt{kin}}$. In the case of the `WWW` both can be approximated by a *discrete* power law.

$$p_{\texttt{kout}} \sim k^{-\alpha_{\texttt{kout}}} \qquad \text{and} \qquad p_{\texttt{kin}} \sim k^{-\alpha_{\texttt{kin}}}.$$

Borrowing the terminology from a branch of statistical physics called the *theory of phase transition*, network whose degree distribution follows a power law are usually called **scale–free**, simply because, as we have seen talking about the Pareto, for specific parameter configurations the selected node's degree could be tiny or arbitrarily large and far away from the mean (exploding variance); that is, without a meaningful internal *scale*.



Now, the next obvious question is: why are hubs and power laws absent in random networks? Or more directly, is there a pool of simple basic rules leading to a Web's growth model with the observed/desired power law behavior? In summary, the crucial observations are the following:

Growth: Real networks are the result of a growth process that continuously increases N. In contrast the random network model assumes that the number of nodes, N, is fixed.

Preferential Attachment: In real networks new nodes tend to link to the more connected nodes. In contrast nodes in random networks randomly choose their interaction partners.

To describe **preferential attachment**, let us work with a very simple model of the `WWW`. The `WWW` consists of web pages and directed hyperlinks from one page to another. The `WWW` is a graph, with pages corresponding to vertices and hyperlinks corresponding to directed edges. The graph grows and changes as pages and links are added to the Web.

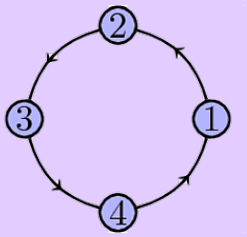
Our model of the Web's growth will be very basic; our goal is not detailed accuracy, but a high–level understanding of what might be happening. Let us start with 2 pages, each linking to the other; the starting configuration does not make a substantial difference, so this configuration is chosen for convenience. At each time step, a new page appears, with just a single link. (One could try to be more accurate by having multiple links or a distribution on links, but having a single link per page simplifies the analysis and yields the important insights).

How should we model what page the new link points to?

The idea behind preferential attachment is that new links will tend to attach to popular pages. In the case of the Web graph, new links tend to go to pages that already have links. We can model this by thinking of the new page as copying a random link, with some probability. Specifically, with probability $\gamma < 1$, the link for the new page points to a page chosen uniformly at random, but with probability $1-\gamma$, the new page copies a random link, so that the new page points to an existing page chosen proportionally to the indegree of that page. We point out that this preferential attachment model of the `WWW` is a Markov chain, as we do not care about the history of how links attached when a new link is added. We only care about the number of links directed into each page. This feature actually simplifies the study of the network key characteristics from a theoreticaly (martingale based) point of view... but this is *not* our goal here.

⇝ Your job ⇜

1. If you haven't already, take a look at basic tools to deal with graphs in `R` such as the `igraph`, `ggraph` packages.
2. Write a program in `R` to simulate the preferential attachment process, starting with 4 pages linked together as a directed cycle on 4 vertices, adding pages each with one outlink until there are 1 million pages, and using $\gamma = 0.5$ as the probability a link is to a page chosen uniformly at random and $1 - \gamma = 0.5$ as the probability a link is copied from existing links.



Simulate a small number `M` of networks and draw a plot of their empirical degree distribution, showing the number of vertices of each degree on a log–log plot. Also draw a plot showing the complimentary cumulative degree distribution – that is, the number of vertices with degree *at least* k for every value of k – on a log–log plot.

Does the degree distribution appear to follow a power law or a Poisson? Explain and comment by showing suitable visual and numerical evidence that supports your reasoning.

Dig a bit more visually and numerically into some of the networks you generated using the `R` tools metioned in (1.)
