Modelling of degree distribution

Thomas Boughen

Abstract Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse aliquam ipsum id consequat euismod. In auctor faucibus mi, vitae blandit nunc vestibulum eget. Fusce eu venenatis dolor, a blandit metus. Aliquam congue congue molestie. Donec euismod bibendum felis a interdum. Vestibulum et euismod sapien, sed accumsan augue. Fusce vulputate arcu et dolor condimentum, a condimentum elit blandit. Nam in maximus lorem. Ut neque felis, hendrerit at orci non, sodales eleifend augue. Duis convallis lacinia justo. Suspendisse imperdiet fermentum finibus. Praesent semper dignissim felis, eu lobortis ex pharetra eu. Pellentesque nec bibendum augue. Cras vulputate viverra finibus. Praesent facilisis quam vel nisi convallis congue. Nunc ullamcorper tristique sagittis.

Data

Here we will be looking at the degree distribution of one specific network, the network of dependencies of packages from the Comprehensive R Archive Network (CRAN). This data forms a network in that, each of the packages has a node and there exists a directed edge between node A and node B if package B depends on package A. We will only be considering the in-degree of nodes of this network as it is more analagous to other real world networks e.g. social media followers.

However this data has a significant number of zeroes that are mostly uninteresting, but will skew the data enough to effect the fit of the models. So, we will be ignoring them from now on, and if we wanted to model them we would model them separately.

Power Law Model

We first consider the model that is often assumed for degree distributions in order to show how well (or poorly) it fits the data. This model, describes the data using the probability mass function:

$$f(x) = \zeta(\alpha + 1)^{-1} x^{-(\alpha+1)} \qquad x = 1, 2, \dots \quad \text{and } \alpha > 0$$

It is possible to fit this model using maximum likelihood, however since we will be using a Bayesian methodology for the other models, it is prudent we do so here as well. First we need the likelihood for a set of data $\boldsymbol{x} = (x_1, x_2, \dots, x_N)^T$

$$L(\boldsymbol{x}) = \zeta(\alpha + 1)^{-N} \prod_{i=1}^{N} x_i^{-(\alpha+1)}$$

Then we need to decide on a prior for α, since it is restricted to the positive real line, we will use $\alpha \sim Ga(1, \lambda_\alpha)$. With this, we can now calculate the posterior distribution :

$$\pi(\alpha|\boldsymbol{x}) = \zeta(\alpha + 1)^{-N} \prod_{i=1}^{N} x_i^{-(\alpha+1)} \lambda_\alpha e^{-\alpha\lambda_\alpha}$$

What is a network?

A network is a collection of nodes with edges between them, these edges can be directed or undirected. The nodes and edges can represent a vast number of things, but representing them in this way allows us to analyse the structure of what is being represented more easily. The degree of a node is the number of edges that are connected to it, with the in-degree and out-degree being the number going in and out respectively. An example is below:
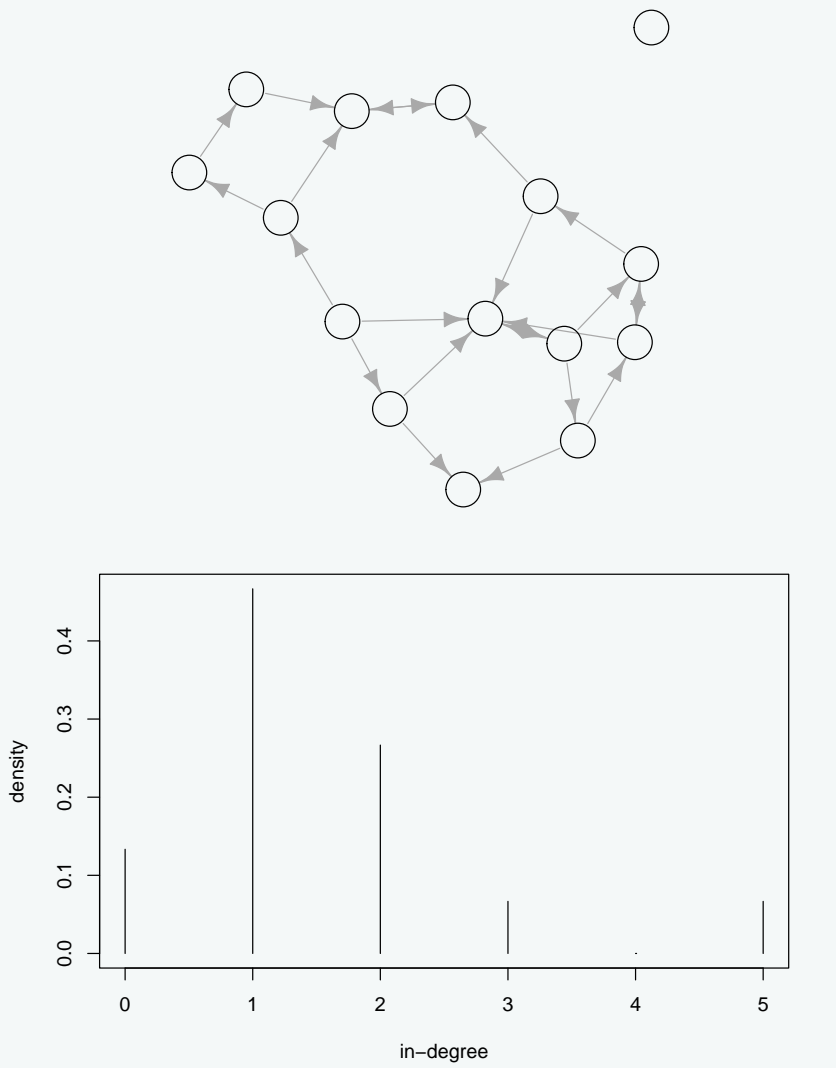
Power Law Mixture

Based on [ref plot] it is seems natural to consider using a mixture of two power laws to model the data. This model assigns the following probability mass function to the data:

$$f(x) = \begin{cases} (1 - \phi)[\zeta(\alpha + 1) - \zeta_v(\alpha + 1)]^{-1} x^{-(\alpha+1)} & , x = 1, 2, \dots, v \\ \phi \zeta_v(\beta + 1)^{-1} x^{-(\beta+1)} & , x = v+1, v+2, \dots \end{cases}$$

where $\alpha, \beta \in \mathbb{R}^+, \phi \in (0, 1), v = \lceil u \rceil$ or $\lfloor u \rfloor$, $u \in \mathbb{R}^+$ and $\zeta_v(s) = \sum_{k=v+1}^{\infty} k^{-s}$.

We decide to use the maximum likelihood estimator of ϕ, $\hat{\phi} = \frac{N-n}{N}$, and decide to use u as an integer between 0 and $N-1$ since in this network no node can have in-degree more than $N-1$.

The priors for the unknown parameters are below:

$$\alpha \sim Ga(1, \lambda_\alpha) \qquad \beta \sim Ga(1, \lambda_\beta) \qquad u \sim U(0, N-1)$$

We can now obtain the posterior distribution:

$$\pi(\alpha, \beta, u|\boldsymbol{x}) = \frac{\lambda_\alpha \lambda_\beta n^n (N-n)^{N-n}}{N^{N+1}} \zeta_u(\beta + 1)^{n-N} [\zeta(\alpha + 1) - \zeta_v(\alpha + 1)]^{-n} \prod_{i=1}^{N} x_i^{-(\gamma_i+1)}$$

where $\gamma_i = \alpha$ if $x_i \leq u$ or $\gamma_i = \beta$ if $x_i > u$, and $n = $ number of $x_i \leq u$.

Power Law-Integrated Generalised Pareto Mixture Model

We naturally look to extreme value theory, specifically using a threshold model. The distribution typically used to describe the distribution of excesses of a threshold is the Generalised Pareto Distribution, but this is only defined for continuous data so we look to using the Integrated Generalised Pareto Distrubtion [CITE] with threshold $v \in \mathbb{Z}^+$, which has probability mass function:

$$g(y) = \left(1 + \frac{\xi(y - v - 1)}{\sigma_0 + \xi v}\right)_+^{-1/\xi} - \left(1 + \frac{\xi(y - v)}{\sigma_0 + \xi v}\right)_+^{-1/\xi}, \qquad y = v+1, v+2, \dots$$

where $\xi \in \mathbb{R}, \sigma_0 \in \mathbb{R}^+$.

Our mixture distribution then has probability mass function:

$$f(x) = \begin{cases} (1 - \phi)[\zeta(\alpha + 1) - \zeta_v(\alpha + 1)]^{-1} x^{-(\alpha+1)} & , x = 1, 2, \dots, v \\ \phi g(x) & , x = v+1, v+2, \dots \end{cases}$$

where $\alpha \in \mathbb{R}^+, \phi \in (0, 1)$ and $\zeta_v(s) = \sum_{k=v+1}^{\infty} k^{-s}$.

We can set $v = \lfloor u \rfloor$ or $\lceil u \rceil$, $u \in \mathbb{R}^+$, here we will set $u \in \mathbb{Z}^+$ and $v = u$

We decide on the priors below for the parameters and use the MLE ($\hat{\phi} = \frac{N-n}{N}$):

$$\alpha \sim Ga(1, \lambda_\alpha), \ \sigma_0 \sim Ga(1, \lambda_\sigma), \ \xi \sim N(0, \lambda_\xi^2), \ u \sim U(0, N-1)$$

and obtain the posterior:

$$\pi(\alpha, \sigma_0, \xi|\boldsymbol{x}) = \frac{\lambda_\alpha \lambda_\sigma n^n (N-n)^{N-n} e^{-(\alpha\lambda_\alpha + \sigma_0\lambda_\sigma + \frac{\xi^2}{2\lambda_\sigma^2})}}{N^{N+1}[\zeta(\alpha + 1) - \zeta_v(\alpha + 1)]^n} \prod_{i:x_i \leq v} x_i^{-(\alpha+1)} \prod_{i:x_i > v} g(x_i)$$

Plots Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse aliquam ipsum id consequat euismod. In auctor faucibus mi, vitae blandit nunc vestibulum eget. Fusce eu venenatis dolor, a blandit metus. Aliquam congue congue molestie. Donec euismod bibendum felis a interdum. Vestibulum et euismod sapien, sed accumsan augue. Fusce vulputate arcu et dolor condimentum, a condimentum elit blandit. Nam in maximus lorem. Ut neque felis, hendrerit at orci non, sodales eleifend augue.
Duis convallis lacinia justo. Suspendisse imperdiet fermentum finibus. Praesent semper dignissim felis, eu lobortis ex pharetra eu. Pellentesque nec bibendum augue. Cras vulputate viverra finibus. Praesent facilisis quam vel nisi convallis congue. Duis convallis lacinia justo. Suspendisse imperdiet fermentum finibus. Praesent semper dignissim felis, eu lobortis ex pharetra eu. Pellentesque nec bibendum augue. Cras vulputate viverra finibus. Praesent facilisis quam vel nisi convallis congue. Duis convallis lacinia justo. Suspendisse imperdiet fermentum finibus. Praesent semper dignissim felis, eu lobortis ex pharetra eu. Pellentesque nec bibendum augue. Cras vulputate viverra finibus. Praesent facilisis quam vel nisi convallis congue.Duis convallis lacinia justo. Suspendisse imperdiet fermentum finibus. Praesent semper dignissim felis, eu lobortis ex pharetra eu. Pellentesque nec bibendum augue. Cras vulputate viverra finibus. Praesent facilisis quam vel nisi convallis congue.

Future Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse aliquam ipsum id consequat euismod. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse aliquam ipsum id consequat euismod. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse aliquam ipsum id consequat euismod.