Scale-freeness and Growth Stability of Realistic Network Models

Project Approval

Thomas Boughen

1 Research Proposal

1.1 Background

Networks often arise when representing complex systems and the relationships between the components within them. Take a social network such as X (formerly Twitter) as an example, new users can join at any time and "follow" other users so that they can keep up with their posts. This forms a directed network where each user corresponds to a node and there exists an edge going from one users node to another if the former "follows" the latter. The number of "followers" of any one user is therefore the in-degree of the corresponding node in the network, is a quantity of interest when it comes to analysing how the network and each node's in-degree may grow in the future. This is just one example of a network, others include:

- Interactions between proteins in cells

- Game results between tennis players

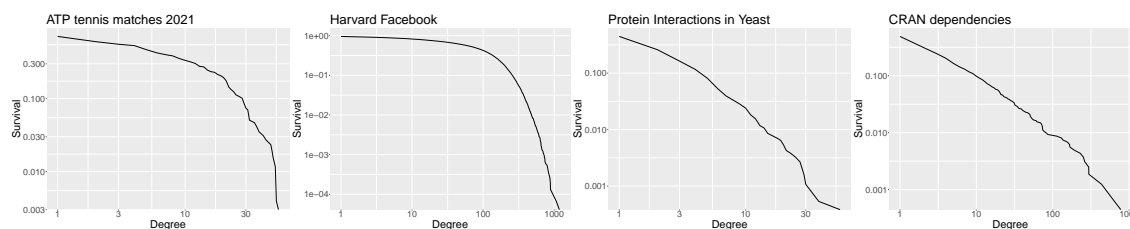- Dependencies of R packages

Figure 1: Survival function of the degrees

The variety of fields across which networks appear make them a vital source of data to be able to analyse, there is particular interest in the way that the networks might have grown. Network generative models describe the way in which nodes and edges are added and removed from networks over time. One might naively propose that nodes in a network are connected to other nodes independently with equal probability, this gives rise to the Random Network Model (Gilbert 1959) which has been proven to generate a network that has a binomial degree distribution which looks nothing like the degree distribution of a lot of real networks. This isn't the only property where this model does not accurately represent real networks. One attempt to create a more accurate network generative model comes from (Barabási and Albert 1999) who noted that the number of nodes in a network grows over time and that nodes tend to connect to the other nodes in the network with more connections (preferential attachment).

1.1.1 Barabási-Albert model

They proposed a minimal model called the Barabási-Albert model, also known as the BA model, which is defined below:

Start with m_0 nodes , the edges between them chosen arbitrarily such that each node has at least one link. The network then develops using two steps:

1. Growth

   At each time step, add a node to the network that will connect to $m \leq m_0$ nodes (already in the network) with m edges.

2. Preferential Attachment

   The probability that an edge from the new node connects to node i is proportional to its current degree k_i i.e $k_i / \sum_j k_j$.

This model generates a network with an approximate power law degree distribution, which looks a lot closer to the degree distributions of real networks and is very simple. Therefore when analysing a network it is tempting to say that it has a power law degree distribution so the BA model can be used to model its growth. The survival function of the degrees of a network produced following this scheme is shown in Figure 2.
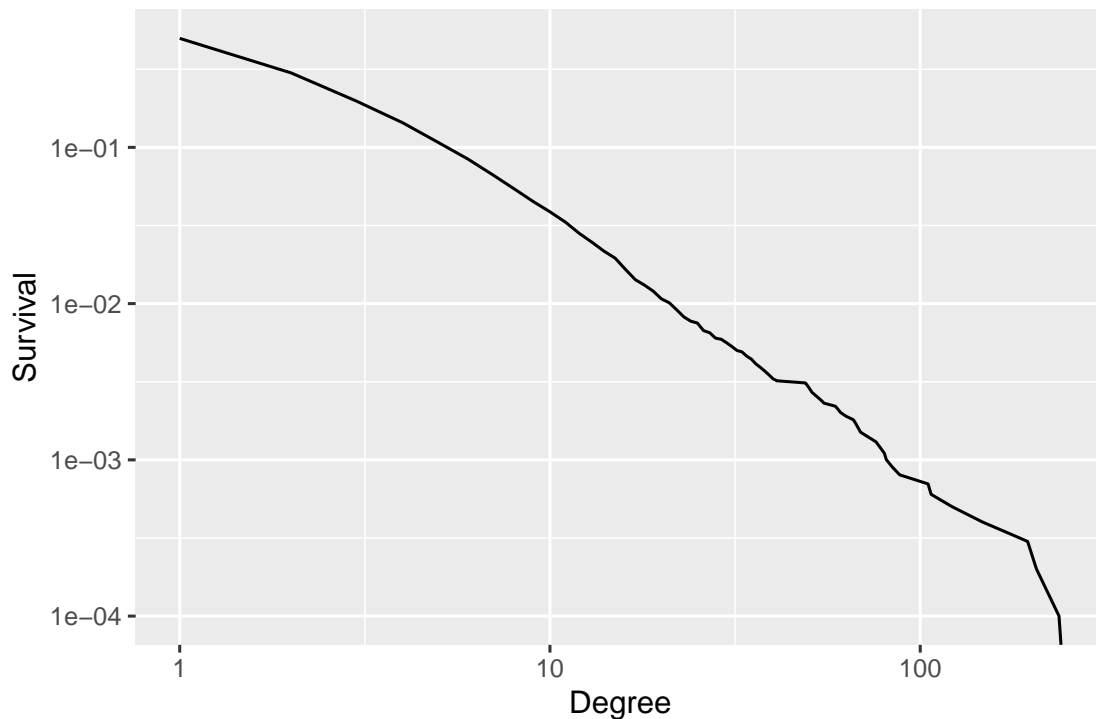

Figure 2: Survival function of the degrees

This does look quite similar to some of the plots in Figure 1 like the CRAN data and protein interactions, but looks very different to the some of the others.

According to recent works (Broido and Clauset 2019; Voitalov et al. 2019) real networks have more nuanced degree distributions and the power law, while suitable for the bulk of the data, is inadequate for the full set of degrees. This is especially clear in the right tail and, to a lesser extent, the left tail. Despite this, methods from extreme value theory remain underutilised even though they seem well suited to the problem of modelling the tails of the degree distribution.

1.2 Aims

This brings us to the aims of this project, which is to firstly find a suitable model for the degree distribution of real networks using methods from extreme value theory, such as a threshold model where the majority of the degrees are modeled by power laws and the right tail is modeled using a discretised version of the Generalised Pareto distribution, the Integer Generalised Pareto distribution (IGP). Then after finding a suitable model for the degree distribution, measuring the changes in model parameters over time for real networks. Informing the modification and/or extension of the BA model to generate networks with the desired degree distribution, such that the new model satisfactorily explains how real networks might have grown. The theoretical properties of networks generated by this model can then be studied, for example how close to following a power law is the networks degree distribution and how do the largest degrees in the network change over time.

1.3 Methodology

As mentioned in the previous subsection, we begin by trying to find a suitable model for the degree distribution of real networks.

1.3.1 Power Law Model

We first look at the power law distribution and fit it to the degree distributions of some real networks using a Bayesian approach.

The power law model is characterised by only one parameter $\alpha \in \mathbb{R}^+$ and has probability mass function (PMF):

$$f(x) = \frac{x^{-(\alpha+1)}}{\zeta(\alpha+1)}, \qquad x = 1, 2, \dots$$

where $\zeta(\alpha+1) = \sum_{k=1}^{\infty} k^{-(\alpha+1)}$ is the Riemann Zeta function.

A good visual tool for judging the fit of the model is the survival function which we define below:

$$S(x) = 1 - \frac{\sum_{k=1}^{x} k^{-(\alpha+1)}}{\zeta(\alpha+1)}, \qquad x = 1, 2, \dots$$

Bayesian Inferenece for the Power Law model

We aim to fit this model using a Bayesian approach and as such we need to define the likelihood for a set of degrees $\boldsymbol{x} = (x_1, x_2, \dots, x_N)^T$. We introduce the auxilliary variables $\boldsymbol{y} = (y_1, y_2, \dots y_m)^T$ and $\boldsymbol{c} = (c_1, c_2, \dots, c_m)^T$ to be the unique values of $x_i \in \boldsymbol{x}$ and the counts of these unique values in the vector $\boldsymbol{x}$ respectively, with the constraint that $N = \sum_{i=1}^{m} c_i$. This is to align the formation of the likelihood with those of the coming models as it will lead to more efficient computations. That said, we now define the likelihood below:

$$L(\boldsymbol{x}) = L(\boldsymbol{y}, \boldsymbol{c}) = \zeta(\alpha+1)^{-N} \prod_{i=1}^{m} y_i^{-c_i(\alpha+1)}$$

Now that we have the likelihood, in order to continue with the Bayesian method we need to decide on a prior for α, we choose an uninformative prior of $\alpha \sim Ga(1, \lambda_\alpha)$ for an adequately small $\lambda_\alpha \in \mathbb{R}^+$. This gives us the posterior distribution of:

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

The results of fitting this model to various data sets are shown in Section 1.4.

1.3.2 Power Law Mixture

Before we can introduce the power law IGP mixture, we would like a model that we can use to benchmark it against. To that end, we look at a mixture of two power laws which seems natural to use as there does seem to be a change in gradient in some of the plots in Figure 1.

This model uses a right truncated power law with exponent $\alpha + 1$ for the bulk of the data and then a left truncated power law with exponent $\beta + 1$ for the remainder.

The PMF of this model is defined below:

$$f(x) = \begin{cases} (1-\phi)\frac{x^{-(\alpha+1)}}{\sum_{k=1}^{v} k^{-(\alpha+1)}}, & x = 1, 2, \ldots v \\ \phi \frac{x^{-(\beta+1)}}{\zeta(\beta+1,v+1)}, & x = v+1, v+2, \ldots \end{cases}$$

where $v = \lceil u \rceil$ for some $u \in \mathbb{R}^+$, $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^+$, $\phi \in (0,1)$ and $\zeta(\beta+1,v+1) = \sum_{k=v+1}^{\infty} k^{-(\beta+1)}$ is the Hurwitz Zeta function.

The survival function is fairly simple to calculate:

$$S(x) = \begin{cases} (1-\phi)\left[1 - \frac{\sum_{k=1}^{x} k^{-(\alpha+1)}}{\sum_{k=1}^{v} k^{-(\alpha+1)}}\right], & x = 1, 2, \ldots, v \\ \phi\left[1 - \frac{\sum_{k=v+1}^{x} k^{-(\beta+1)}}{\zeta(\beta+1,v+1)}\right], & x = v+1, v+2, \ldots \end{cases}$$

Bayesian Inferenece for the Power Law Mixture Model

We again calculate the likelihood using the auxiliary variables $\boldsymbol{y}$ and $\boldsymbol{c}$:

$$L(\boldsymbol{x}) = (1-\phi)^n \phi^{N-n} \zeta(\beta+1,v+1)^{n-N} \left(\sum_{k=1}^{v} k^{-(\alpha+1)}\right)^{-n} \prod_{i:y_i \leq v} y_i^{-c_i(\alpha+1)} \prod_{i:y_i > v} y_i^{-c_i(\beta+1)}$$

We now specify the priors used to calculate the posterior distribution:

$$\alpha \sim N(0, 1/\lambda_\alpha)$$
$$\beta \sim Ga(1, \lambda_\beta)$$
$$\phi \sim Beta(1, 1)$$

These are all relatively uninformative priors for any adequately small values of λ_α and λ_β. We do not decide on a prior for u directly, instead choosing to specify it through ϕ.

We now continue to calculate the posterior distribution:

$$\pi(\alpha, \beta, u|\boldsymbol{x}) \propto \lambda_\alpha^{-1} \lambda_\beta \exp\{-(\beta\lambda_\beta + \alpha^2\lambda_\alpha)\} \zeta(\beta+1,v+1)^{n-N} \left(\sum_{k=1}^{v} k^{-(\alpha+1)}\right)^{-n}$$
$$\times \prod_{i:y_i \leq v} y_i^{-c_i(\alpha+1)} \prod_{i:y_i > v} y_i^{-c_i(\beta+1)}$$

We fit this model for the same sets of data as we did for the previous models and the results are show next in Section 1.4.

1.3.3 Power Law IGP Mixture Model

To define this model (PL-IGP) we will first need to look at a traditional way of modelling extreme values, more specifically approximating the conditional distribution of the largest values. This is often done using the Generalised Pareto distribution (GPD) is parameter by the scale parameter $\sigma \in \mathbb{R}^+$, the shape parameter $\xi \in \mathbb{R}$, and a threshold $u \in \mathbb{R}^+$. It is denoted by $Y|Y > u \sim GPD(\sigma, \xi)$ where Y is some continuous random variable. The GPD has survival function defined below:

$$\Pr(Y > y|Y > u) = \left(1 + \frac{\xi(y-u)}{\sigma}\right)_+^{-1/\xi}, \qquad y > u$$

where $a_+ = \max(0, a)$.

However, this is only really appropriate to use for approximating the conditional distribution of the largest values of continuous random variables. So, we would like a similar distribution that can be used for discrete random variables.

Integrated Generalised Pareto Distribution (IGPD)

We follow roughly the same method as Rohrbeck et al. (2018), but they only consider modelling $Y = \lceil H \rceil$ we also consider the possibility of modelling $Y = \lfloor H \rfloor$ where H is a continuous random variable with support on the positive real line such the $H|H > u \sim GPD(\sigma_0 + \xi u, \xi)$ for some $u \in \mathbb{R}^+$. Below we define the PMF of the IGPD for both cases.

Ceiling Version: the case where $Y = \lceil H \rceil$

For values of $x = \lfloor u \rfloor, \lfloor u \rfloor + 1, \ldots$ and $\xi \in \mathbb{R}$ and $u, \sigma\_0 \in \mathbb{R}^+$:

$$\Pr(Y = x|Y > u) = \Pr(H < x|H > \lfloor u \rfloor) - \Pr(H < x - 1|H > \lfloor u \rfloor)$$
$$= \left(1 + \frac{\xi(x - \lfloor u \rfloor)}{\sigma_0 + \xi\lfloor u \rfloor}\right)_+^{-1/\xi} - \left(1 + \frac{\xi(x - 1 - \lfloor u \rfloor)}{\sigma_0 + \xi\lfloor u \rfloor}\right)_+^{-1/\xi}$$

Floor Version: the case where $Y = \lfloor H \rfloor$

For values of $x = \lceil u \rceil, \lceil u \rceil + 1, \ldots$ and $\xi \in \mathbb{R}$ and $u, \sigma_0 \in \mathbb{R}^+$:

$$\Pr(Y = x|Y > u) = \Pr(H < x + 1|H > \lceil u \rceil) - \Pr(H < x|H > \lceil u \rceil)$$
$$= \left(1 + \frac{\xi(x + 1 - \lceil u \rceil)}{\sigma_0 + \xi\lceil u \rceil}\right)_+^{-1/\xi} - \left(1 + \frac{\xi(x - \lceil u \rceil)}{\sigma_0 + \xi\lceil u \rceil}\right)_+^{-1/\xi}$$

For the sake of briefness, from here on out we will always be using the floor version of the IGPD, but note that the results can be easily extended for the ceiling version.

This now allows us to define the mixture model of a truncated power law and the IGPD, which has the PMF:

$$f(x) = \begin{cases} (1-\phi)\frac{x^{-(\alpha+1)}}{\sum_{k=1}^{v} k^{-(\alpha+1)}}, & x = 1, 2, \ldots, v \\ \phi\left[\left(1 + \frac{\xi(x+1-v)}{\sigma_0+\xi v}\right)_+^{-1/\xi} - \left(1 + \frac{\xi(x-v)}{\sigma_0+\xi v}\right)_+^{-1/\xi}\right], & x = v+1, v+2, \ldots \end{cases}$$

For $\phi \in (0, 1)$, $\sigma_0 \in \mathbb{R}^+$, $\alpha \in \mathbb{R}$, $\xi \in (-\sigma_0/v, \infty)$, and $v = \lceil u \rceil$ for $u \in \mathbb{R}^+$.

To simplify the expressions that follow we define:

$$G(x) = \left(1 + \frac{\xi(x-v)}{\sigma_0 + \xi v}\right)_+^{-1/\xi}$$

Again, we want the survival function so we can get an idea of how well the model performs for real data. It is defined as:

$$S(x) = \begin{cases} (1-\phi)\left[1 - \frac{\sum_{k=1}^{x} k^{-(\alpha+1)}}{\sum_{k=1}^{v} k^{-(\alpha+1)}}\right], & x = 1, 2, \ldots, v \\ \phi G(x+1), & x = v+1, v+2, \ldots \end{cases}$$

We continue to fit this model using a Bayesian methodology, so we define the likelihood again using the auxiliary variables $\boldsymbol{y}$ and $\boldsymbol{c}$ and we obtain the below:

$$L(\boldsymbol{x}) = L(\boldsymbol{y}, \boldsymbol{c}) = \phi^{N-n}(1-\phi)^n \left(\sum_{k=1}^{v} k^{-(\alpha+1)}\right)^{-n} \prod_{i:y_i \leq v} y_i^{-c_i(\alpha+1)} \prod_{i:y_i > v} [G(y_i+1) - G(y_i)]^{c_i}$$

where n is the number of values in $\boldsymbol{x}$ that are at most v.

Bayesian Inferenece for the PL-IGP model

Proceeding with the Bayesian approach, we decide on the following priors for the parameters:

$$\alpha \sim N(0, 1/\lambda_\alpha),$$
$$\xi \sim N(0, 1/\lambda_\xi),$$
$$\sigma_0 \sim Ga(1, \lambda_\sigma),$$
$$\phi \sim Beta(1, 1)$$

Again, we are using fairly uninformative priors for the parameters. Using the same method of indirectly specifying a prior for u through the prior of ϕ instead.

With all of this we find the posterior distribution to be:

$$\pi(\alpha, \sigma_0, \xi, u | \boldsymbol{x}) \propto (2\pi)^{-1} \lambda_\alpha^{1/2} \lambda_\xi^{1/2} \lambda_\sigma \exp\{-[\sigma_0 \lambda_\sigma + \frac{1}{2}(\alpha^2 \lambda_\alpha + \xi^2 \lambda_\xi)]\} \left(\sum_{k=1}^{v} k^{-(\alpha+1)}\right)^{-n}$$
$$\times \prod_{i:y_i \leq v} y_i^{-c_i(\alpha+1)} \prod_{i:y_i > v} [G(y_i+1) - G(y_i)]^{c_i}$$

As before, the results of fitting this model to several data sets are found in Section 1.4.

1.4   Work So Far

So far, we have started looking for a model that is adequate for modelling the degree distributions of real networks. We have used some adaptive Metropolis-Hastings algorithms using methods similar to those in *Handbook of Markov Chain Monte Carlo* (2011) and Xiang and Neal (2014) to obtain estimates and credible intervals for the model parameters. Below we show the results of this for the same data sets shown in Figure 1.

1.4.1 The Power Law

In Figure 3 below we can see that for some sets of data the power law does an okay job for the bulk of the data, but it doesn't fit well in the right tail usually lying above the empirical survival curve. For other sets of data, the model seems completely inadequate namely for the Tennis and Harvard data sets. These results were obtained using the prior parameter $\lambda_\alpha = 0.01$ and an initial value of $\alpha = 1$ in an adaptive Metropolis-Hastings algorithm for 100,000 iterations with a burn in period of 500 iterations and then thinned by 5, yielding a minimum effective sample size of 2,000.
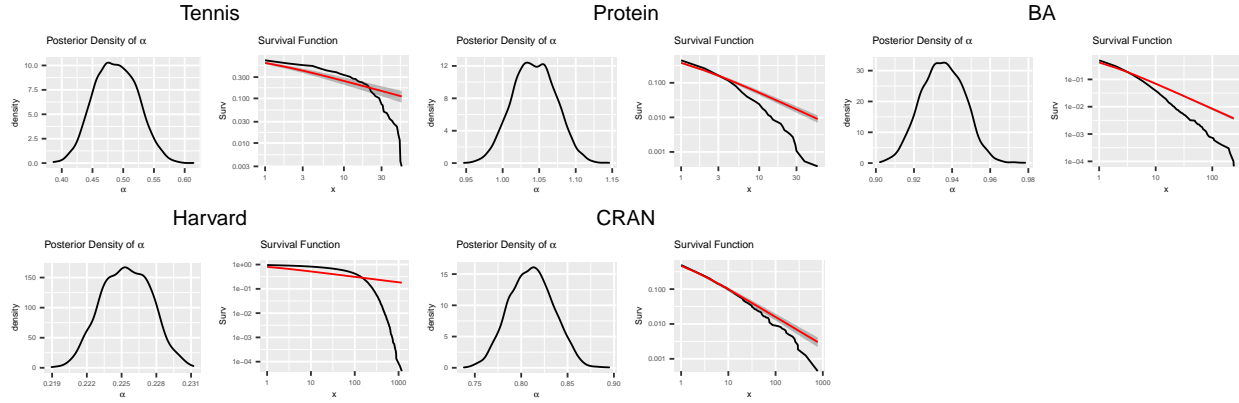

Figure 3: Plots of fitted survival functions and posterior densities of α

1.4.2 Power Law IGP Model

This mixture model defined in Section 1.3.3 seems to perform significantly better for all of the data sets. Within each group of plots in Figure 4, the posterior densities are shown for each of the parameters and then in the bottom left of each group is the 95% credible interval for the survival. While this model seems to fit these sets of data well, it may be needlessly complex and have more parameters than necessary. So, next we fit the model described in Section 1.3.2.
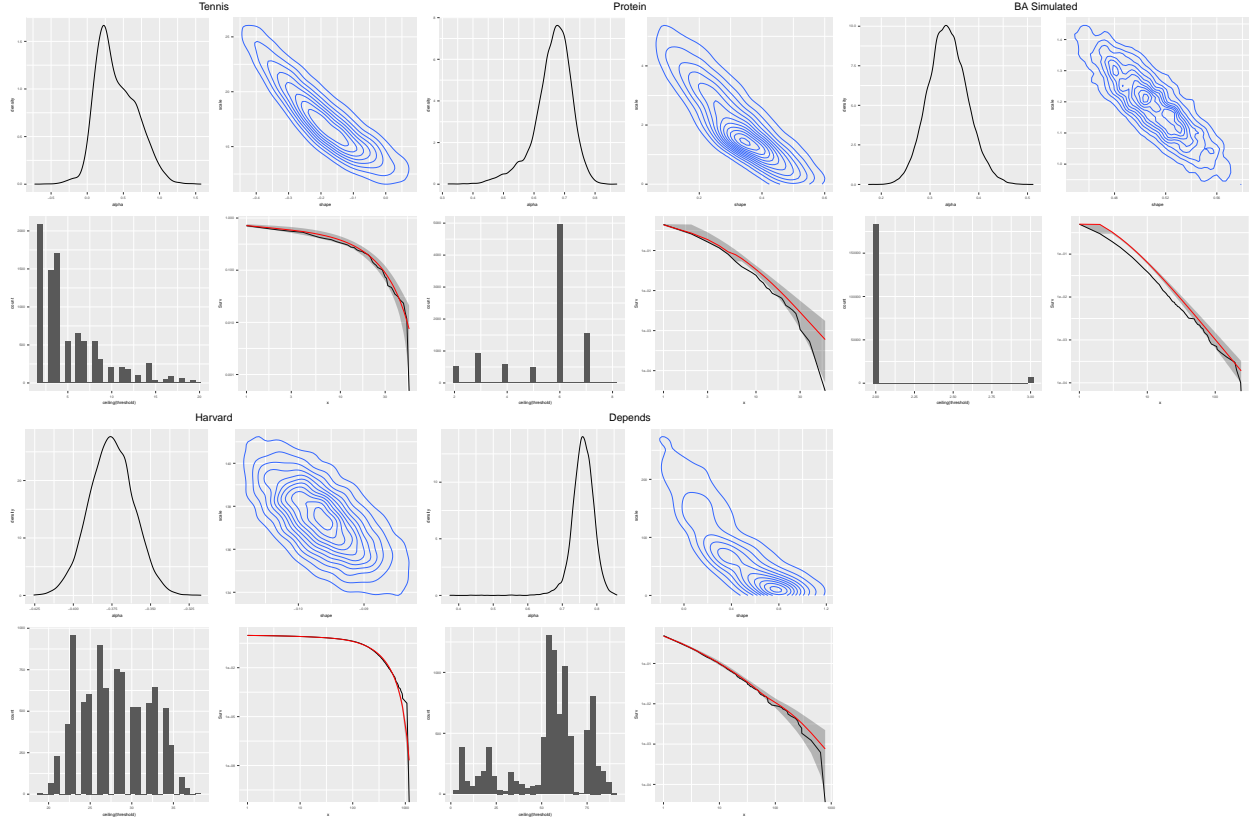
Figure 4: Plots of fitted survival functions and posterior densities for the parameters

1.4.3 Power Law Mixture

Looking at Figure 5, this model also seems to perform quite well for all of the sets of data. In the top left of each group of plots is a 2D density plot of the two different power law parameters, α and β. In the top right is a histogram of the value of $\lceil u \rceil$ and finally, in the bottom right is a plot of the empirical survival function with the posterior mean in red and 95% credible intervals for the survival. It is interesting to note that ,for the simulated data, the value of $\lceil u \rceil$ did not stray from the value of 3 indicating that using only one power law may be suitable which was expected. This also happens to be the minimum value that the code will allow the threshold to take since model selection has not been implemented.
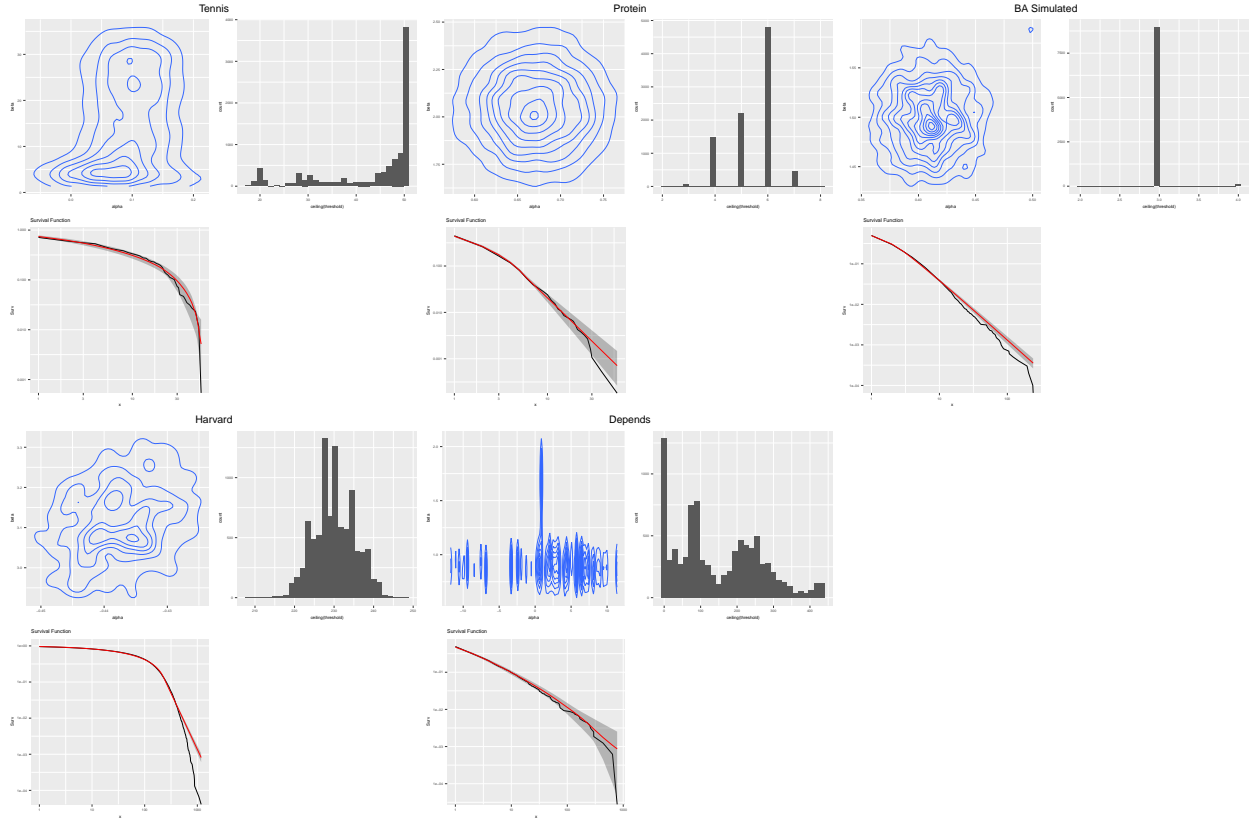
Figure 5: Fitted survival functions and posterior densities of parameters

1.4.4 Summary

So far we have shown that the power law is only adequate for certain sets of data, like the BA model and CRAN data, and seems to not be useful for modelling many other sets of data. Introducing the power law IGP mixture improves the performance of the model across almost all of the data sets but is a lot more complex. To compromise between the two we also included a power law mixture model, which seems to have yielded similar results to the power law IGP mixture while having less parameters. This will need to be investigated in more depth because it may end up being preferable due to it being less complex.

Upon further investigation and seeing how well these models fit to various sets of data, we will either decide to use one of them, or decide to look for an alternative model to use. Then we can begin moving along two directions, in one we look at tracking the changes in parameters of the chosen model over time for various real data sets as well as data simulated from the BA model. This informs what we do on the other, where we will be attempting to modify and/or extend the BA model so that the new model will provide a degree distribution that follows the chosen model and matches up with the changes in parameters over time for the data sets.

2 Project Plan

Review proofs of properties of the Barabási-Albert model

This will involve mostly reading of the original papers detailing the properties of the BA model, with some attempt to recreate the proofs perhaps in a more mathematically rigorous fashion. This may also help in future when it comes to deriving the properties of the new model.

Read required extreme value theory and network science materials

This will also involve mostly reading of materials that will help in understanding the mathematics behind network generative models in addition to some extreme value theory that will be needed when it comes to properly defining the model we choose for degree distributions and its properties.

Review previous modifications of the Barabási-Albert model

Before we can attempt to make any modifications to the BA model, we will need to review what extensions/modifications have already been done.

Come up with a suitable model for the degree distribution

Building on the previous steps, this is where we will choose a model to use for the degree distribution so that we can begin making appropriate modifications to the BA model in the next steps.

Create set of functions and/or R package to fit the model

After coming up with a suitable model, we will require a way to fit it in a somewhat uninvolved process so that the model can be fit to many sets of data over many time frames.

Monitor changes in model parameters for various real networks

This will be an ongoing process throughout some of the other steps, all while using the results produced to inform how we carry out the other steps.

Use these changes to inform modification of the BA model

We will begin investigating possible modifications to the BA model that will produce a degree distribution that aligns with the chosen model and hopefully with the changes in the model parameters over time as the networks grow.

Investigate properties of the new model

Once a set of modifications has been made, we will then study the theoretical properties of networks that grow under the new model.

References

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5439): 509–12. https://doi.org/10.1126/science.286.5439.509.

Broido, Anna D., and Aaron Clauset. 2019. "Scale-Free Networks Are Rare." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-08746-5.

Gilbert, E. N. 1959. "Random Graphs." *The Annals of Mathematical Statistics* 30 (4): 1141–44. https://doi.org/10.1214/aoms/1177706098.

Handbook of Markov Chain Monte Carlo. 2011. Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

Rohrbeck, Christian, Emma F. Eastoe, Arnoldo Frigessi, and Jonathan A. Tawn. 2018. "Extreme value modelling of water-related insurance claims." *The Annals of Applied Statistics* 12 (1): 246–82. https://doi.org/10.1214/17-AOAS1081.

Voitalov, Ivan, Pim van der Hoorn, Remco Hofstad, and Dmitri Krioukov. 2019. "Scale-Free Networks Well Done." *Physical Review Research* 1 (October). https://doi.org/10.1103/PhysRevResearch.1.033034.

Xiang, Fei, and Peter Neal. 2014. "Efficient MCMC for Temporal Epidemics via Parameter Reduction." *Computational Statistics & Data Analysis* 80: 240–50. https://doi.org/https://doi.org/10.1016/j.csda.2014.07.002.