Annual Progress Review

Thomas William Boughen

May 31, 2024

Newcastle University

School of Mathematics, Statistics and Physics

[1]An html version of this report can be found at www.twboughen.github.io/phd/APR/report

Table of contents

1. Introduction **3**

2. Extreme Value Theory **4**
2.1. Continuous Extremes 4
2.2. Discrete Extremes 6
2.3. Modelling 7

3. Networks **8**
3.1. Mathematical Definitions 8
3.2. Network Generative Models 8

4. Methods **11**
4.1. Modelling degree distributions 11
4.2. Fitting model to the data 12
4.3. GPA analyses 15
4.4. A Conjecture 16

5. Next Steps **17**

A. Updated Project Plan **18**

B. Training **19**
Funding and Stipend 19

References **20**

1. Introduction

Since the aim is to gain understanding about the behaviour of the degree distribution of networks at the right tail, it seems natural to look to using methods from extreme value theory.

2. Extreme Value Theory

This section begins with a review of the theory and methodology for modelling the extreme values of continuous random variables, before moving to considerations for modelling the extreme values of discrete random variables.

2.1. Continuous Extremes

Studying the properties of the extreme values of a random variable first requires determining what exactly is considered to be an extreme value. In this section extreme values of two kinds are considered, both of which can be characterised.

The first kind of extreme value considers the distribution of block maxima. That is, for a set of independent and identically distributed (iid) random variables $X_1, \dots, X_n$ with common cumulative density function (cdf) F what is the limiting distribution of $M_n = \max\{X_1, \dots, X_n\}$?

Clearly, as $n \to \infty$, the block maxima M_n converges almost surely to the right endpoint of F. However, standardising the block maxima allows for some characterisation of the limiting distribution.

Theorem 2.1.1 (Fisher–Tippett–Gnedenko Theorem)**.** *With $X_1, \dots, X_n \overset{\text{iid}}{\sim} F$ and $\{a_n\}_{n\geq 0}, \{b_n\}_{n\geq 0}$ such that:*

$$\lim_{n\to\infty} \Pr\left(\frac{1}{a_n}[M_n - b_n] \leq x\right) = G(x),$$

for some non-degenerate G.

Then F is said to be in the (maximum) domain of attraction of G, denoted $F \in \mathcal{D}(G)$,and G is of one of three types:

- *Gumbel:* $\Lambda(x) = \exp\{-\exp(-x)\}, \quad x \in \mathbb{R}$
- *Fréchet:* $\Phi_\alpha(x) = \exp\{-x^{-\alpha}\}, \quad x \geq 0, \alpha > 0$
- *Negative-Weibull:* $\Psi_\alpha(x) = \exp\{-x^{-a}\}, \quad x < 0, \alpha > 0$

Each of these three types defines a domain of attraction.

Definition 2.1.1 (Domains of Attraction)**.** The three domains of attraction that result from Theorem 2.1.1 have the following equivalent conditions:

For a distribution with cdf F and survival function $\bar{F}$ that has right endpoint x_F given by:

$$x_F = \sup\{x \in \mathbb{R} \cup \{\infty\} : F(x) < 1\}$$

the distribution belongs to each domain of attraction subject to the conditions below:

If there exists a positive function a

- Type I/Gumbel/$\mathcal{D}(\Lambda)$:

$$\lim_{x\uparrow x_F} \frac{\bar{F}(x + ta(x))}{\bar{F}(x)} = e^{-t}, \quad \forall t \in \mathbb{R}$$

If $x_F = \infty$:

- Type II/Fréchet/$\mathcal{D}(\Phi_\alpha)$:

$$\lim_{x\to\infty}\frac{\bar{F}(tx)}{\bar{F}(x)} = x^{-\alpha}, \quad \forall t>0 \quad \text{for some } \alpha>0$$

If $x_F < \infty$:

- Type III/Negative-Weibull/$\mathcal{D}(\Psi_\alpha)$:

$$\lim_{h\downarrow 0}\frac{\bar{F}(x_F - xh)}{\bar{F}(x_F - h)} = x^{\alpha}, \quad \alpha>0$$

The parameter α in Definition 2.1.1 and Theorem 2.1.1 is called the extreme value index.

Here, distributions in the Gumbel domain are referred to as light tailed, distributions in the Negative-Weibull domain are referred to as short tailed, and those in the Fréchet are referred to as heavy tailed.This terminology for heavy tailed distributions in different ot some of the literature that defined a heavy tailed distribution as one that decays slower than exponential. However the terminology used here is also widely used.

Throughout this report functions will be referred to as regularly varying or slowly varying, what is meant by this is formally deined below:

Definition 2.1.2 (Regular Variation)**.** A positive,real valued, measurable function f is said to be regularly varying at infinity with index γ if for all $t>0$:

$$\lim_{x\to\infty}\frac{f(tx)}{f(x)} = x^{\gamma}.$$

If $\gamma = 0$, then f is instead said to be slowly varying at infinity.

Note that the condition for a distribution to belong to the Fréchet domain of attraction is equivalent to saying that the survival function $\bar{F}$ is regularly varying with index $-\alpha$.

In addition to heavy tailed distributions it is also useful to define what will be referred to as super heavy tailed distributions. This term is often just refers to specific distributions such as the log-Cauchy ,log-Gamma,and log-Weibull distributions but Fraga Alves, Haan, and Neves (2009) provides a more precise definition below:

Definition 2.1.3 (Super Heavy Tails)**.** A distribution is with survival function $\bar{F}$ is said to have super heavy tails if:

$$\lim_{x\to\infty}\frac{\bar{F}(tx)}{\bar{F}(x)} = 1, \qquad \forall t>0$$

That is, a distribution is called super heavy if its survival function is slowly varying.

The three main types of extremal distribution (Gumbel, Fréchet and Negative-Weibull) can be united into one distribution, called the Generalised Extreme Value (GEV) distribution.

Definition 2.1.4 (Generalised Extreme Value Distribution)**.** Denoted by $\text{GEV}(\mu,\sigma,\xi)$ the distribution is characterised by three parameters $\mu\in\mathbb{R}$ the location, $\sigma\in\mathbb{R}^+$ the scale, and the shape $\xi\in\mathbb{R}$. It has support on $\{x\in\mathbb{R} : 1+\xi(x-\mu)/\sigma>0\}$ and has cdf given by:

$$G(x) = \begin{cases} \exp\left\{-\left(1+\dfrac{\xi(x-\mu)}{\sigma}\right)_+^{-1/\xi}\right\}, & \xi\neq 0 \\ \exp\left\{-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right\}, & \xi = 0. \end{cases}$$

The three types of extremal distribution are obtained from changing the shape parameter ξ, which corresponds to $1/\alpha$ in Theorem 2.1.1. This change is generally made so that the largest ξ corresponds to heavier tails of the distribution. Specifically, $\xi<0$, $\xi=0$, $\xi>0$, correspond to the negative Weibull, Gumbel and the Fréchet domains of attraction respectively.

Another kind of extreme values are the observations above a large threshold, like the limiting distribution of block maxima, the limiting distribution of these extreme values can be characterised by the generalised pareto (GP) distribution.

Definition 2.1.5 (Generalised Pareto Distribution)**.** Consider a random variable X with the same cdf F as in Theorem 2.1.1, the Generalised Pareto (GP) distribution can be obtained by using the GEV distribution and conditional probability such that for large enough threshold the GP distribution approximately describes the conditional distribution of threshold exceedances. More precisely, for sufficiently large threshold u and the change of variable to $Y = X - u$:

$$\Pr(Y \leq y | Y > 0) = H(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi y}{\sigma}\right)^{-1/\xi}, & y > 0, \xi \neq 0 \\ 1 - \exp\left(-\dfrac{y}{\sigma}\right), & y > 0, \xi = 0 \end{cases}$$

Since this distribution was obtained using a GEV(μ, σ^*, ξ) the shape parameter ξ is identical in both distributions and the shape parameter σ is defined such that $\sigma = \sigma^* + \xi(u - \mu)$.

It is also possible to derive the result without using the GEV, as shown in [REF].

2.2. Discrete Extremes

A lot of Section 2.1 is appropriate only for continuous random variables and some of the results may not hold in a discrete setting. In particular, a continuous distribution F being in certain domain of attraction may not necessarily imply that a discretisation of F remains in that domain of attraction.

Definition 2.2.1 (Discretisation)**.** The discretisation of a distribution with cdf F is given by

$$F^*(n) = F(n) - F(n-1), \quad n \in \mathbb{Z}$$

Shimura (2012) provides conditions for a discretisation of a continuous distribution to belong to the same domain of attraction. In particular the following theorem which corresponds to Theorem 1 in Shimura (2012).

Theorem 2.2.1 (Domain of attraction consistency)**.**

(a) Every discretisation of distribution in $\mathcal{D}(\Phi_\alpha)$ remains in $\mathcal{D}(\Phi_\alpha)$.
(b) The discretisation of a distribution remains in $\mathcal{D}(\Lambda)$ if and only if the original is in $\mathcal{D}(\Lambda) \cap \mathcal{L}$.

Where $\mathcal{L}$ is the set of long-tailed distributions that have the property:

$$\lim_{x \to \infty} \frac{\overline{F}(x+1)}{\overline{F}(x)} = 1$$

In addition Shimura (2012) introduces a quantity useful for determining the domain of a attraction that a discrete distribution belongs to.

Definition 2.2.2 (Omega Function)**.** For a distribution F with survival function $\overline{F}$ and some $n \in \mathbb{Z}^+$ let:

$$\Omega(F, n) = \left(\log \frac{\overline{F}(n+1)}{\overline{F}(n+2)}\right)^{-1} - \left(\log \frac{\overline{F}(n)}{\overline{F}(n+1)}\right)^{-1}$$

This quantity plays an important role in Section 4 when determining the domain of attraction to which the degree distribution of a network generative model belongs. In particular a discrete distribution is recoverable to the Fréchet domain of attraction $\mathcal{D}(\Phi_\alpha)$ if:

$$\lim_{n \to \infty} \Omega(F, n) = \alpha^{-1}$$

Applying ideas from Section 2.1 to modelling discrete random variables has been approached from many different directions. What follows is a overview of some of the approaches that have been taken but will see use in this report.

Hitz, Davis, and Samorodnitsky (2024) note that using the GP distribution as an approximation in a discrete setting leads to bias in the likelihood function and can lead to it being inadequate for modelling. They propose

two other peaks over threshold methods that rely on parametric families of discrete distributions. The first, what they refer to as the discrete generalised Pareto approximation is based on an extension of the discrete survival function. The second, the generalised Zipf distribution is obtained from an extension of the probability mass function. Both methods are motivated theoretically for modelling of a large class of discrete distributions and are shown in the paper to either match or outperform using the GP to model discrete data directly.

Ahmad, Gaetan, and Naveau (2022) first introduce an extended GP distribution, a continuous distribution that extends the idea of obtaining GP values from a probability integral transform (PIT) of $U(0,1)$ draws and instead considers a PIT of draws from any distribution on $(0,1)$ such as a beta distribution. This distribution is then discretised into their discrete extended GP distribution.

2.3. Modelling

The results from Section 2.1 allow the GEV and GP to be fitted to the block maxima and exceedances respectively. Typically, when fitting the GP, a sufficiently high threshold needs to be specified beforehand. [COLES] provides some empirical methods for specifying the threshold, one approach is to use a threshold stability plot that uses maximum likelihood to estimate the parameters of the GP for a large range of thresholds. The threshold can be chosen as the point across all of the plots after which the values of the parameters seems stable.

Another more recent approach shown in [MACDONALD 2012], uses a spliced threshold mixture to model the threshold exceedances where one distribution is assumed for the bulk of the data and the GP is used for those values above the threshold. This approach can also be applied in the discrete setting, and is what is used in Section 4.

3. Networks

Networks are the data sources that the results from Section 2 will be used to analyse. Networks appear across a wide range of fields when attempting to represent complex systems and the relationships between the components within them.

This section will being with an introduction to the basics of networks and working with them in mathematics and probability, including the concept of degree distribution. Then, a look at a few network generation models and limiting results for the degree distributions of the networks they generate.

3.1. Mathematical Definitions

Throughout this section, graphs constructed from vertices and edges will be used as an analogue for these networks, so it is appropriate to begin with some mathematical definitions for exactly what that means.

Definition 3.1.1 (Graph)**.** A graph $G = (V, E)$ is constructed from a vertex set V and an edge set E. The edge set can take on one of two forms depending on if the graph is directed or un-directed. If the graph is directed then $E \subseteq V^2$ i.e the edge set is contained within the set of ordered pairs of vertices, whereas if the graph is **un-directed** then $E \subseteq [V]^2$, where

$$[V]^2 = \{\{u, v\} : u, v \in V\}$$

i.e. the edge set is contained within the set of un-ordered pairs of vertices.

Definition 3.1.2 (Degree of un-directed graphs)**.** For an un-directed graph a vertex's degree denoted $d(v)$ for $v \in V$ is the number of edges that are connected to vertex v:

$$d(v) = |\{e \in E : v \in e\}|$$

Definition 3.1.3 (Degree of directed graphs)**.** Directed graphs have something analogous, called the in-degree d_{in}, out-degree d_{out} and total degree d_{tot}. The in-degree of a vertex v is the number edges with endpoint at v, whereas the out-degree is the number of edges with start point at v and the total degree is the sum of these i.e.:

$$\begin{aligned} d_{in}(v) &= |\{(w_1, w_2) \in E : w_2 = v\}| \\ d_{out}(v) &= |\{(w_1, w_2) \in E : w_1 = v\}| \\ d_{tot}(v) &= d_{in}(v) + d_{out}(v) \end{aligned}$$

There are many reasons to analyse network like data, one of which is to gain an insight into the mechanics that governed the growth of the network. The next sub-section is focused on presenting several network generative models, that may be able to describe how real networks grow. For now, the focus will be on the degree distributions of these network generative models.

3.2. Network Generative Models

It is useful to be able to model the way a network may have grown using simple rules as the subsequent model can then be used to simulate how the network may grow in future and provide insights into the underlying mechanics of the system the network represents. These models are also sometimes called mechanistic models in the literature. Also, although they are referred to as network generative models, graphs are still being used in the rules that govern how the generative model works.

This section begins by detailing a fairly simple generative model and its limiting results for the degree distribution, followed by two special cases of the first model and their results.

General Preferential Attachment (GPA)

Under this model, at each time step one vertex is added to the network and brings an edge with it that connects the existing vertices with a probability proportional to some function of the vertices degrees.

Definition 3.2.1 (General Preferential Attachment Model)**.** Starting with a graph $G_1 = (V_1, E_1) = (\{1\}, \emptyset)$, at each following time step $t > 1$ the graph $G_t = (V_t, E_t)$ is generated by the following rules:

1. **Growth:** Add a new vertex to the vertex set i.e.

$$V_t = V_{t-1} \cup \{t\}$$

2. **Preferential Attachment:** Add one edge connecting the new vertex to one already in the graph $\{1, \dots, t-1\}$ selected at random with weights proportional to a function of their degree i.e.:

$$E_t = E_{t-1} \cup \{\tilde{e}\}$$

where $\tilde{e}_i = \{t, \tilde{v}\}$ and $\tilde{v} = i$ with probability

$$\frac{g(d(i))}{\sum_{w \in V_{t-1}} g(d(i))}, \qquad i \in V_{t-1}$$

for some function $g : \mathbb{Z} \mapsto \mathbb{R}^+ \quad \{0\}$, which will be referred to as the preferential attachment function

Limiting Degree Distribution

In [GPA REF] the limiting degree distribution was calculated in terms of the preferential attachment function and does not have a general explicit form. It is defined as follows, let λ^* be the solution, if it exists, to:

$$1 = \sum_{n=1}^{\infty} \prod_{i=1}^{n-1} \frac{g(i)}{g(i) + \lambda}$$

then the limiting degree distribution of a network resulting from the GPA model has probability mass function (pmf):

$$f(k) = \frac{\lambda^*}{g(k) + \lambda^*} \prod_{i=0}^{k-1} \frac{g(i)}{g(i) + \lambda^*}$$

Barabási-Albert (BA)

The GPA model has several special cases, when g is the identity function i.e $g(k) = k$, it becomes the BA model (Barabási and Albert 1999) with $m = 1$.

Definition 3.2.2 (Barabási-Albert Model)**.** Starting with a graph $G_1 = (V_1, E_1)$ where $V_1 = \{1, \dots, m_0\}$ and $E_1 = \{\{v\} : v \in V_1\}$ i.e a graph with m_0 vertices with one self-loop each. At each time step $t > 1$ the graph $G_t = (V_1, E_1)$ is generated by the following rules:

1. **Growth:** Add a new vertex to the vertex set i.e.

$$V_t = V_{t-1} \cup \{t\}$$

2. **Preferential Attachment:** Add $m \leq m_0$ edges between the new vertex and those already in the graph with probability proportional to each vertices degree i.e.

$$E_t = E_{t-1} \cup \{\tilde{e}_1, \dots, \tilde{e}_m\}$$

where each new edge $\tilde{e}_i = \{t, \tilde{v}_i\} (i = 1, \dots, m)$ has $\tilde{v}_i$ sampled independently without replacement from V_{t-1} with probability:

$$\frac{d(\tilde{v}_i)}{\sum_{u \in V_{t-1}} d(u)}$$

Limiting Degree Distriubtion

In [NETWORK SCI book], it was shown that for large values of t, the limiting degree distribution of a network produces by this model is:

$$f(k) = \frac{2m(m+1)}{k(k+1)(k+2)}, \qquad k \geq m$$

According to [REF] this pmf is regularly varying with exponent 2, meaning that it is in the Fréchet domain of attraction $\mathcal{D}(\Phi_2)$.

Uniform Attachment (UA)

The final special case presented here is obtained from setting the preferential attachment function g to be some constant value.

Definition 3.2.3 (Uniform Attachment Model)**.** Start with a graph $G_1 = (V_1, E_1) = (\{1, \dots, m_0\}, \emptyset)$, at each time step $t > 1$ the graph is denoted by $G_t = (V_t, E_t)$ and generated by repeating the following two steps:

1. **Growth:** Add a new vertex to the vertex set i.e.

$$V_t = V_{t-1} \cup \{t\}$$

2. **Uniform Attachment:** Add $m \leq m_0$ edges between the new vertex and those already in the graph with probability proportional to each vertices degree i.e.

$$E_t = E_{t-1} \cup \{\tilde{e}_1, \dots, \tilde{e}_m\}$$

where each new edge $\tilde{e}_i = \{t, \tilde{v}_i\} (i = 1, \dots, m)$ has $\tilde{v}_i$ sampled independently without replacement from V_{t-1} with probability:

$$\frac{1}{\sum_{u \in V_{t-1}} 1} = \frac{1}{|V_{t-1}|}$$

Limiting Degree Distribution

As showing in [REF] the expected degree distribution of this model for large values of t is approximately:

$$f(k) = \frac{e}{m} \exp\left(-\frac{k}{m}\right), \qquad k \geq m$$

Although this was not shown rigourously and treats the degree of a vertex as a continuous random variable, this is an shifted exponential distribution with left endpoint m and rate parameter $1/m$ and as such is in the Gumbel domain of attraction.

If $m = 1$, it is possible to get a more precise result from the result regarding the limiting degree distribution of the GPA. By setting the preferential attachment function $g(k) = \lambda^*$, the can be shown that the limiting degree distribution is:

$$f(k) = \left(\frac{1}{2}\right)^k, \qquad k = 1, 2, \dots$$

This distribution also occupies the Gumbel domain of attraction.

4. Methods

The aim of this section is to investigate the degree distribution of real networks and compare them to the results obtained for the generative models in Section 3.2. First, a look at what the degree distributions of real networks look like.

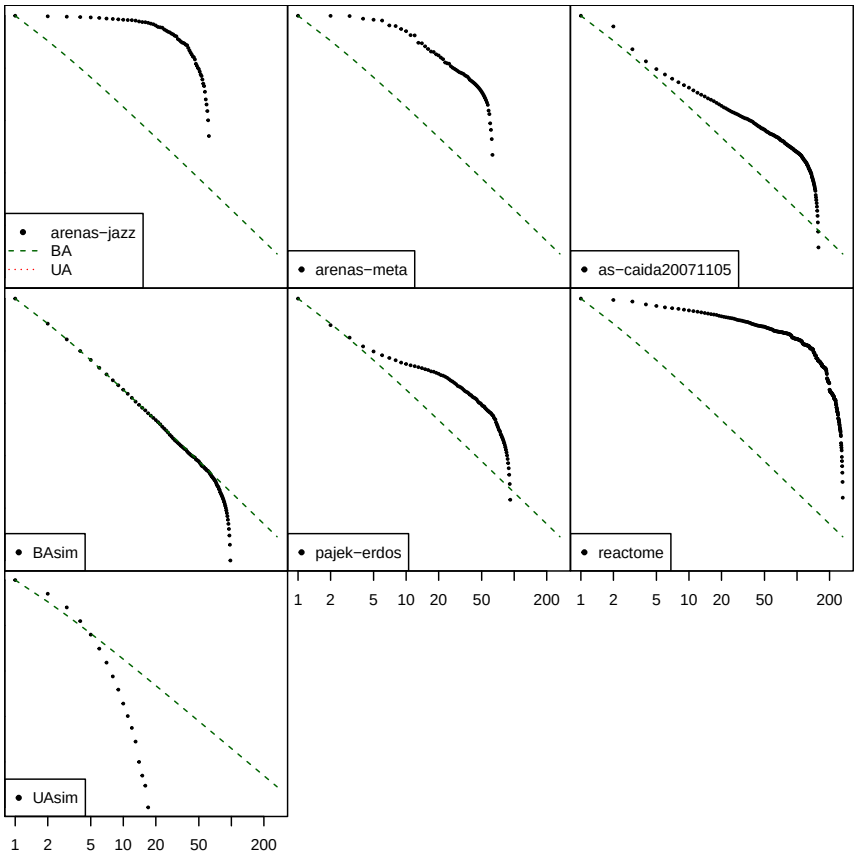

Figure 4.1.: Plots of survial functions of real networks degrees

Figure 4.1 shows the survival function of the degrees of various real networks as well as "BAsim" and "UAsim" which were generated using the corresponding schemes in Section 3.2. Additionally, the theoretical limiting degree distribution of both the UA model and the BA model (for m=1) are included on the plots. Visually it seems that neither of these models are adequate for modelling the growth of the real networks shown here.

To further investigate this, Section 4.1 considers fitting a model to these data that will provide insight into what would be needed from a network generative model such that it flexible enough to capture the variation of shapes of degree distribution in real networks.

4.1. Modelling degree distributions

As mentioned in Section 2.3, the method used here to model the extreme values of the data will be a spliced threshold mixture. Specifically, it will be a spliced threshold mixture of a power law and a discretisation of the generalised pareto distribution similar to what is defined in [ROHRBECK].

Definition 4.1.1 (Intergral Generalised Pareto Distribution (IGP))**.** Consider a random variable X with cdf F, and consider the random variable $Y = \lfloor X \rfloor$. From Definition 2.1.5, $X|X > u \sim GP(\sigma, \xi)$ for some sufficiently large $u \in \mathbb{R}^+$ and it can be obtained that the distribution of $Y|Y > u$ has distribution defined below:

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

For $y = \lceil u \rceil, \lceil u \rceil + 1, ...$ and $\xi \in \mathbb{R}$ and $u, \sigma_0 \in \mathbb{R}^+$.

Since the some degree distributions of real networks seen in Figure 4.1 seem to be approximately linear for the bulk of the data and then begin to change, the spliced threshold mixture that will be used consists of a truncated discrete power law for the bulk of the data and a GP above a threshold.

Definition 4.1.2 (Power-Law IGP Distribution)**.**

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

4.2. Fitting model to the data

The values of the parameters in the model for each data set were estimated under the Bayesian framework using a Metropolis within Gibbs sampler. Below are plots showing the same data as in Figure 4.1 but with the mean and 95% confidence intervals of the survival function of the model for each data-set.

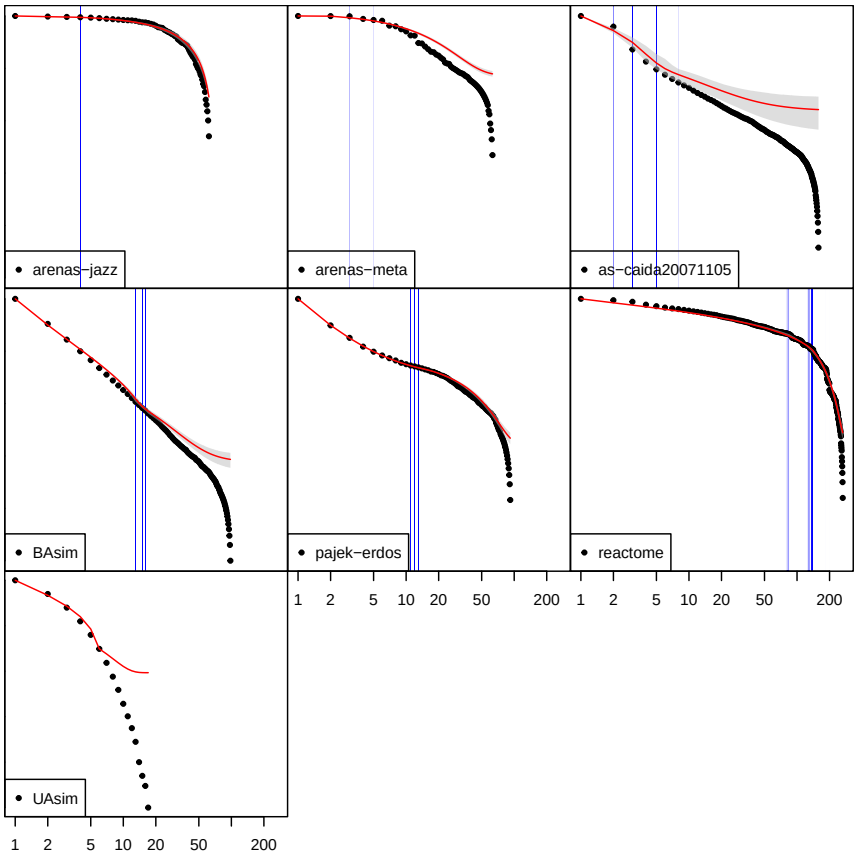

Figure 4.2.: Plots of truncated survial functions of real networks degrees

Unfortunately, this model does not seem sufficient for these data. Perhaps, it may be better to add an additional component to the spliced mixture that is used to model the left tail. Since the focus is on the more extreme values in the distributions, instead of fitting the model to the full set of data, some of the lower values in the data sets will be removed in order to improve model fit for the remaining data. Again, in future this will

likely be replaced by a third component in the spliced mixture. The fits to the truncated data sets are shown below in Figure 4.3.

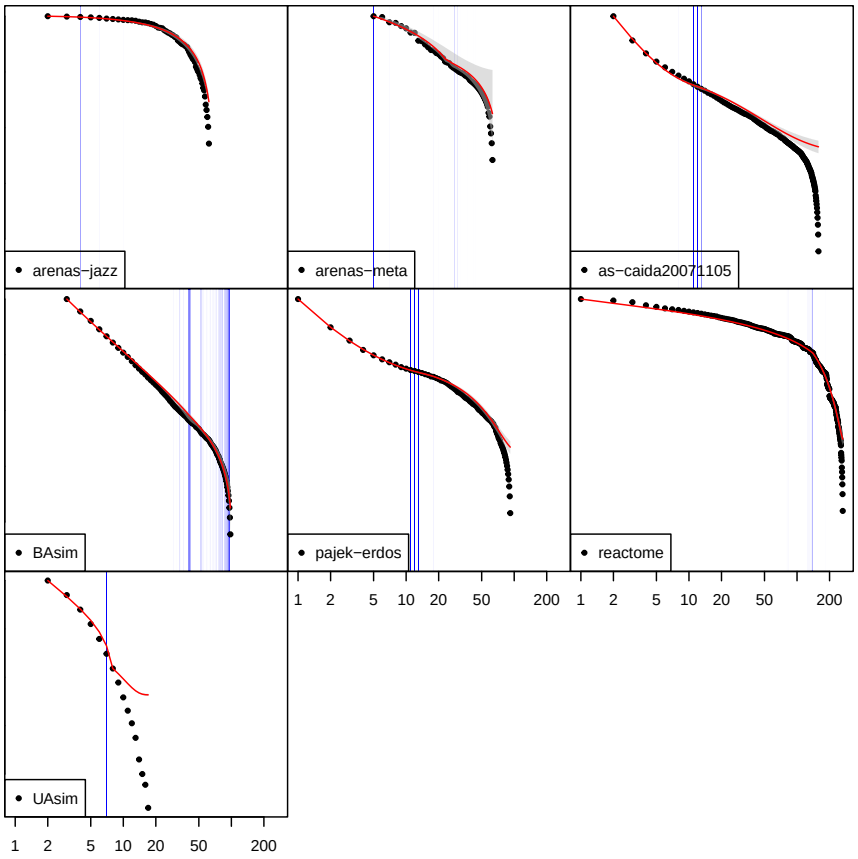

Figure 4.3.: Plots of truncated survial functions of real networks degrees

As show by Figure 4.3 left truncating the data works well to improve the model fit for the vast majority of the data sets. The main exception being the data simulated via the UA model. It is unclear why the model fit for this is so poor but it may be because of the low number of unique values of degrees. That said, below is a summary of the model parameters for each data set:

Table 4.1.: arenas-jazz

phi	a	xi	sig	v
Min. :0.8290	Min. : 0.00333	Min. :-0.9540	Min. :35.04	Min. : 2.000
1st Qu.:0.9845	1st Qu.: 2.44954	1st Qu.:-0.7678	1st Qu.:43.47	1st Qu.: 2.000
Median :0.9845	Median : 28.49493	Median :-0.7199	Median :45.89	Median : 2.000
Mean :0.9809	Mean : 40.82553	Mean :-0.7233	Mean :46.09	Mean : 2.227
3rd Qu.:0.9845	3rd Qu.: 81.00571	3rd Qu.:-0.6742	3rd Qu.:48.72	3rd Qu.: 2.000
Max. :0.9845	Max. :100.00000	Max. :-0.5470	Max. :59.77	Max. :12.000

Table 4.2.: arenas-meta

phi	a	xi	sig	v
Min. :0.002632	Min. :0.001132	Min. :-1.3389	Min. : 5.829	Min. : 5.00
1st Qu.:0.152632	1st Qu.:0.122445	1st Qu.:-0.8916	1st Qu.:44.027	1st Qu.:22.00
Median :0.152632	Median :0.207592	Median :-0.7787	Median :50.714	Median :22.00
Mean :0.221311	Mean :0.344512	Mean :-0.7086	Mean :48.331	Mean :21.44
3rd Qu.:0.152632	3rd Qu.:0.343474	3rd Qu.:-0.6485	3rd Qu.:57.528	3rd Qu.:22.00
Max. :0.965789	Max. :2.059039	Max. : 0.3328	Max. :84.683	Max. :62.00

Table 4.3.: as-caida20071105

phi	a	xi	sig	v
Min. :0.03477	Min. :1.916	Min. :0.1950	Min. : 6.248	Min. : 7.00
1st Qu.:0.05424	1st Qu.:2.077	1st Qu.:0.3563	1st Qu.:10.151	1st Qu.:10.00
Median :0.05424	Median :2.100	Median :0.3981	Median :11.496	Median :11.00
Mean :0.05532	Mean :2.100	Mean :0.3989	Mean :11.527	Mean :10.93
3rd Qu.:0.06016	3rd Qu.:2.125	3rd Qu.:0.4405	3rd Qu.:12.774	3rd Qu.:11.00
Max. :0.08604	Max. :2.338	Max. :0.6145	Max. :19.167	Max. :17.00

Table 4.4.: BAsim

phi	a	xi	sig	v
Min. :6.012e-05	Min. :1.562	Min. :-4.3169	Min. : 0.4234	Min. :29.00
1st Qu.:1.202e-04	1st Qu.:1.613	1st Qu.:-1.2654	1st Qu.: 51.9085	1st Qu.:52.00
Median :5.411e-04	Median :1.623	Median :-0.6266	Median : 67.1963	Median :89.00
Mean :2.096e-03	Mean :1.623	Mean :-0.5006	Mean : 87.0407	Mean :76.73
3rd Qu.:3.908e-03	3rd Qu.:1.634	3rd Qu.:-0.4248	3rd Qu.:125.9865	3rd Qu.:96.00
Max. :1.184e-02	Max. :1.676	Max. :29.8810	Max. :490.8584	Max. :97.00

Table 4.5.: pajek-erdos

phi	a	xi	sig	v
Min. :0.03667	Min. :1.269	Min. :-0.43807	Min. :19.68	Min. : 9.00
1st Qu.:0.04331	1st Qu.:1.330	1st Qu.:-0.26402	1st Qu.:28.44	1st Qu.:12.00
Median :0.04706	Median :1.346	Median :-0.21989	Median :30.50	Median :13.00
Mean :0.04655	Mean :1.346	Mean :-0.22031	Mean :30.57	Mean :13.33
3rd Qu.:0.04908	3rd Qu.:1.361	3rd Qu.:-0.17868	3rd Qu.:32.47	3rd Qu.:15.00
Max. :0.05645	Max. :1.440	Max. : 0.02085	Max. :44.27	Max. :19.00

Table 4.6.: reactome

phi	a	xi	sig	v
Min. :0.04457	Min. :9.940e-07	Min. :-0.4912	Min. : 62.10	Min. : 80.0
1st Qu.:0.10684	1st Qu.:1.763e-04	1st Qu.:-0.3301	1st Qu.: 84.46	1st Qu.:140.0
Median :0.10684	Median :4.007e-04	Median :-0.3002	Median : 90.63	Median :140.0
Mean :0.11069	Mean :4.775e-04	Mean :-0.3017	Mean : 91.48	Mean :137.8
3rd Qu.:0.10684	3rd Qu.:4.763e-04	3rd Qu.:-0.2684	3rd Qu.: 96.60	3rd Qu.:140.0
Max. :0.21558	Max. :4.340e-03	Max. :-0.1467	Max. :132.28	Max. :184.0

Table 4.7.: UAsim

phi	a	xi	sig	v
Min. :0.01518	Min. :1.281	Min. :-0.3375	Min. :4.346	Min. :7
1st Qu.:0.01518	1st Qu.:1.310	1st Qu.:-0.3011	1st Qu.:4.992	1st Qu.:7
Median :0.01518	Median :1.317	Median :-0.2899	Median :5.156	Median :7
Mean :0.01518	Mean :1.317	Mean :-0.2904	Mean :5.160	Mean :7
3rd Qu.:0.01518	3rd Qu.:1.325	3rd Qu.:-0.2797	3rd Qu.:5.300	3rd Qu.:7
Max. :0.01518	Max. :1.356	Max. :-0.2283	Max. :5.794	Max. :7

4.3. GPA analyses

So far it has been shown that neither the BA model nor the UA model can adequately capture the range of type of degree distributions of real networks. So, a natural place to start when attempting to expand the range of possible degree distributions is the more general model, the GPA. This section, will use results from Shimura (2012) and Section 2 to investigate the possible types of degree distribution that may arise from different preferential attachment functions in the GPA model.

The Preferential Attachment Function

From here on the preferential functions that will be used for the GPA model will be of the form:

$$g(k) = k^{\gamma}, \qquad \gamma > 0.$$

This allows for investigating the cases where the preferential attachment function is sub-linear and when it is super-linear.

As discussed in Section 2.2, the limiting value of $\Omega(F, n)$ can give a lot of information about the behaviour of a discrete distribution at extreme values. Below is a plot showing the value of this quantity as n increases for various different values of γ.

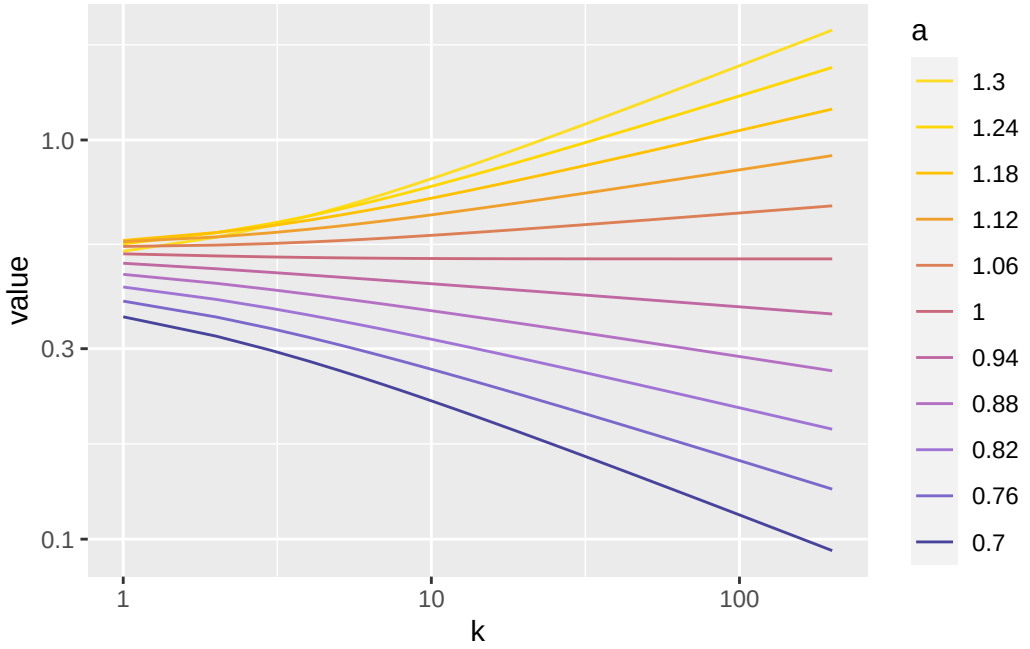

Figure 4.4.: Plot of $\Omega(F, n)$ for various $\gamma \in (0.7, 1.3)$

Figure 4.4 shows that for $\gamma < 1$ $\Omega(F, n)$ seems to approach 0 as n increases, whereas for $\gamma = 1$ $\Omega(F, n)$ seems to converge to finite non-zero limit which is to be expected as this corresponds to the BA model which has limiting degree distribution in the Fréchet domain of attraction. However, for $\gamma > 1$ the value of $\Omega(F, n)$ appears to diverge and does not approach a finite limit.

Shimura (2012) does not provide any results in particular for the case of $\Omega(F, n)$ diverging but if the definition of slow variation and thus super-heavy tails is viewed as regular variation in the limit as α goes to infinity then the following can be obtained.

Corollary 4.3.1. *For a distribution F with survival function $\overline{F}$ and some $n \in \mathbb{Z}^+$, if:*

$$\lim_{n\to\infty} \Omega(F, n) = \lim_{\alpha\downarrow 0} \alpha^{-1} = \infty$$

then F has super heavy tails

This is further supported by Figure 4.5 below, which shows the value of the quantity from Definition 2.1.3 for increasing values of n and values of γ in the range $(1, 2)$. The plot shows the quantity approaching 1 for all values of γ as n increases, suggesting that the limiting degree distribution of the GPA model with $g(k) = k^\gamma, \gamma > 1$ has super heavy tails.

4.4. A Conjecture

The results from this subsection suggest that for super-linear preferential attachment functions the GPA model has limiting degree distribution with super heavy tails. This, along with results for the linear case in Section 3.2 and sub-linear cases in [NETSCI BOOK] lead to the following conjecture.

Conjecture 4.4.1. *The GPA model is only capable of producing three different types of degree distribution:*

1. *Gumbel: sub-linear preferential attachment function*
2. *Fréchet $\mathcal{D}(\Phi_2)$: linear preferential attachment function*
3. *Super heavy tails: super-linear preferential attachment function*

This means that under the framework presented here, even the GPA model is no where near close to being able to capture the range of types of degree distribution found in real networks.

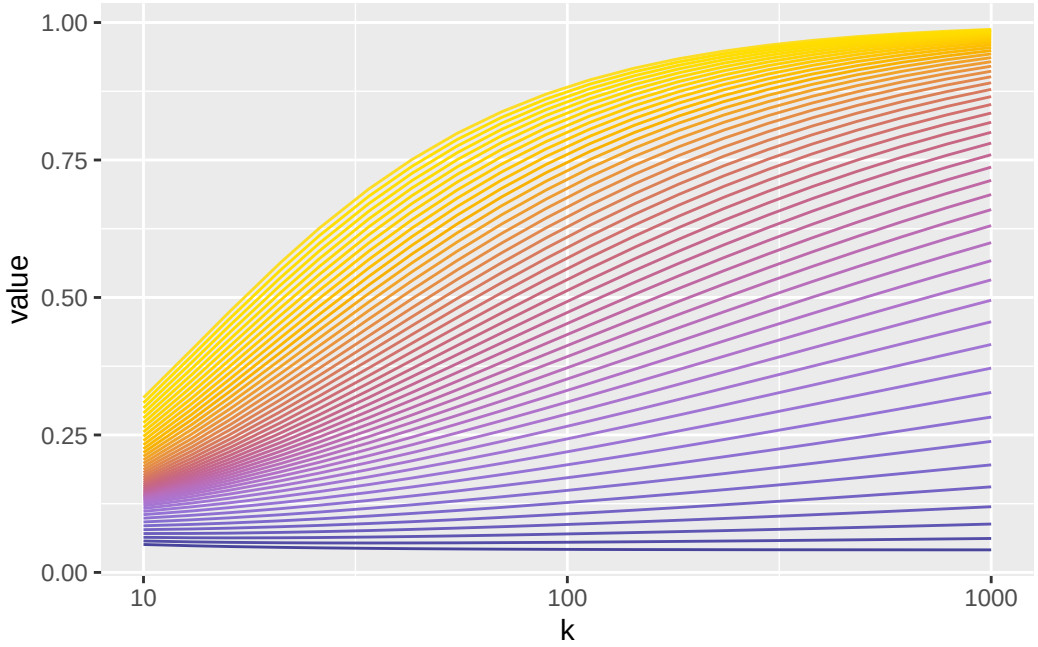

Figure 4.5.: Plot testing slow variation for $\gamma \in (1, 2)$

5. Next Steps

A. Updated Project Plan

B. Training

Funding and Stipend

The funding for this project expires on **17th March 2027**.

References

Ahmad, Touqeer, Carlo Gaetan, and Philippe Naveau. 2022. "Modelling of Discrete Extremes Through Extended Versions of Discrete Generalized Pareto Distribution." *ArXiv e-Prints*. https://arxiv.org/abs/2210.15253.

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.

Fraga Alves, Maria, Laurens Haan, and Cláudia Neves. 2009. "A Test Procedure for Detecting Super-Heavy Tails." *Journal of Statistical Planning and Inference* 139 (February). https://doi.org/10.1016/j.jspi.2008.04.026.

Hitz, Adrien S., Richard A. Davis, and Gennady Samorodnitsky. 2024. "Discrete Extremes." *Journal of Data Science*, 1–13. https://doi.org/10.6339/24-JDS1120.

Shimura, Takaaki. 2012. "Discretization of Distributions in the Maximum Domain of Attraction." *Extremes* 15: 299–317. https://doi.org/10.1007/s10687-011-0137-7.