Modeling the GPRS network latency with a double Pareto-lognormal or a generalized Beta distribution

Bernd Pfitzinger[1], Tommy Baumann[2], Andreas Emde[2], Daragan Macos[3], and Thomas Jestädt[1]

[1]*Toll Collect GmbH, Linkstraße 4, 10785 Berlin, Germany*
[2]*Andato GmbH & Co. KG, Ehrenbergstraße 11,98693 Ilmenau, Germany*
[3]*Beuth Hochschule für Technik, Luxemburger Straße 10, 13353 Berlin, Germany*

Abstract

Taking a newly collected large data set on the TCP connection termination latency in GPRS networks we try to identify the underlying statistical distribution. The data extends the observed latencies to large time scales necessitating a heavy-tail distribution. Many distributions work well for the main body of the data. However, the heavy tail of the distribution benefits from mixing different statistical distributions. We compare several distributions and find that the double Pareto-lognormal distribution and the generalized Beta distribution of the second kind fit the data equally well.

1. Introduction

Many phenomena – technical or natural – exhibit rare events at a much higher rate than the normal distribution: At least one of the tails of the probability distribution follows a power-law. Examples range from personal income [1], insurance claims, wildfires, gene mutation or network size [2] (see e. g. [3] for empirical examples and [4] for a historical perspective).

We add the network connection termination latency to these examples. We have collected a large data set of time intervals between two successive server-side log entries in the German automatic toll system. The log entries correspond to the successful acknowledgment of the toll data received by the central server followed by the successful termination of the TCP connection (see figure 1). Without further access to the system we assume that at least a total of four TCP packets – one pair each to close the two-way connection from each end – are exchanged between the onboard unit (OBU) and the central server. All OBU types are equipped with 2G GSM modems up to class 10.

We recorded a total of more than 300 million data points for more than 1 000 000 OBUs operating within the reach of one of the German mobile networks. To our knowledge the largest data set in the literature contains 12 million data points [5], collected in seven (unnamed) countries by passive monitoring of the connection setup. While passive monitoring – i.e. with access to the core network – has many advantages, the connection start terminates by default after three seconds thereby severely curtailing the observation of rare events.

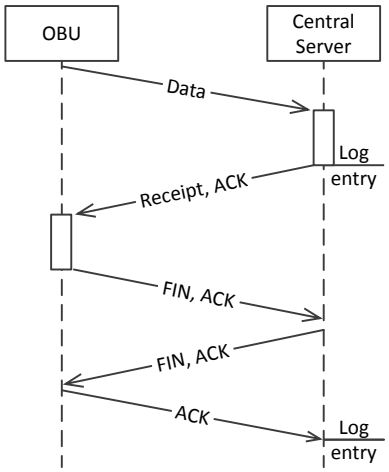

Figure 1: Sequence diagram of the TCP/IP communication between an on-board unit (OBU) and the central system.

In this article we take the collected data and try to identify statistical distributions that are able to describe the observed behavior. To that extent we limit ourselves to the overall data set, i.e. all events regardless of the OBU type or mobile network. We note in passing that the data exhibits several interesting features (e.g. peaks in the probability density function corresponding to the retransmission of one packet or even the exponential

URI: http://hdl.handle.net/10125/50619
ISBN: 978-0-9981331-1-9

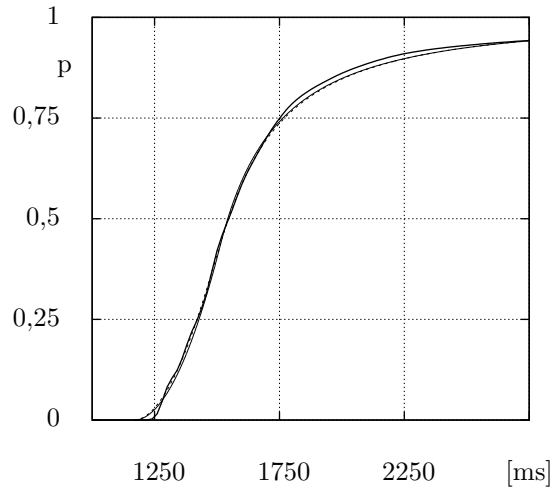

Figure 2: Cumulative distribution function of the connection termination latency. Observed data (bold line) compared with two statistical distributions: Double Pareto-lognormal (dashed) and generalized Beta distribution of the second kind (thin line).

back-off algorithm) which are not the topic of this article.

Figure 2 summarizes the data set as the cumulative distribution function (CDF) of all successful connection termination events (bold line in figure 2). Some connections will not terminate successfully e.g. only after reaching a timeout – the most pronounced occurring after about 70 seconds. In our analysis we disregard these (few) connections and scale the data accordingly (i.e. the CDF of the successful connection termination events reaches up to 99.68 %).

The next sections introduce two different statistical distributions that we find to describe the data: The double Pareto-lognormal distribution in the next section and the generalized Beta distribution of the second kind thereafter. The technical implementation is almost trivial with some notable exceptions mentioned in section 4. The results are summarized in section 5.

2. Double Pareto lognormal distribution

The Pareto distribution is the typical power-law distribution but lacking the ability to model the body of an empirical distribution. Many possible distributions are listed in the literature (e.g. [6] summarizes the application of several distributions to empirical data). For practical purposes we choose only distributions having a closed form for the PDF and CDF preferentially using only readily available functions. For the Pareto distribution this leads naturally to the double Pareto-lognormal distribution.

The double Pareto-lognormal distribution (dPlN, [7], [8]) is a mixture of two heavy-tail Pareto distributions and a lognormal distribution modeling the body of the distribution. Consequently it requires four parameters – the power-law scaling of the two Pareto-like heavy-tails (α, β) and the mean and variance of the lognormal distribution (ν, τ^2).

From a practical point of view the dPlN distribution has the considerable advantage of having a closed form for the probability density function (PDF) f_{dPlN} (equation 1) and the CDF F_{dPlN} (equation 3) depending on four parameters (α, β for the power-law tail behavior and ν, τ for the body of the distribution, $x > 0$):

$$f_{\text{dPlN}}(x,\alpha,\beta,\nu,\tau) = \frac{\alpha\beta}{\alpha+\beta}\cdot \left(x^{-\alpha-1} A(\alpha,\nu,\tau)\Phi\left(\frac{\log x - \nu - \alpha\tau^2}{\tau}\right) + x^{\beta-1} A(-\beta,\nu,\tau)\Phi^c\left(\frac{\log x - \nu + \beta\tau^2}{\tau}\right)\right) \quad (1)$$

$$\text{with } A(\alpha,\nu,\tau) = e^{\alpha\nu+\alpha^2\tau^2/2}. \quad (2)$$

The equations are taken from the online version of [7] and corrected for typographical and arithmetic errors as noted in the appendix of [9]. A correct derivation of the CDF is given in [10] albeit with the wrong sign in the final result. The CDF F_{dPlN} should read:

$$F_{\text{dPlN}}(x,\alpha,\beta,\nu,\tau) = \Phi\left(\frac{\log x - \nu}{\tau}\right) - \frac{\beta x^{-\alpha}}{\alpha+\beta} A(\alpha,\nu,\tau)\Phi\left(\frac{\log x - \nu - \alpha\tau^2}{\tau}\right) + \frac{\alpha x^{\beta}}{\alpha+\beta} A(-\beta,\nu,\tau)\Phi^c\left(\frac{\log x - \nu + \beta\tau^2}{\tau}\right) \quad (3)$$

where Φ is the CDF of the normal distribution:

$$\Phi(x) = \frac{1}{2}\Big(\text{erf}(x/\sqrt{2}) + 1\Big) \text{ and } \Phi^c(x) = 1 - \Phi(x).$$

3. Generalized Beta distribution of the second kind

While the dPlN-distribution is often mentioned as the better choice its direct competitor seems to be the much older generalized Beta distribution of the second kind (GB2): First developed in the 1980s [11] (for recent summaries see [12], [13] or [14]) it also uses four parameters (α, β, p, q all of which we choose to be positive) and produces a power-law tail.

The GB2-distribution is a generalization that reduces for specific parameters to a number of well-known distributions: The beta distribution of the second kind, the generalized gamma distribution, two different Burr type distributions, the lognormal, Weibull, gamma, Lomax, F, Fisk, Rayleigh, χ^2, half-normal, half-Student's t, exponential and log-logistic distributions ([11], [15]). Equation 4 lists the PDF f_{GB2} of the GB2-distribution for $x > 0$ depending on the Beta function $\mathrm{B}(p, q)$:

$$f_{\text{GB2}}(x, p, q, \alpha, \beta) = \frac{|\alpha|}{\beta \mathrm{B}(p, q)} \left(\frac{x}{\beta}\right)^{\alpha p - 1} \left(1 + (x/\beta)^{\alpha}\right)^{-p-q}. \quad (4)$$

The CDF F_{GB2} also exists in a closed form, depending amongst others on the incomplete Beta function $I_x(a, b)$ which is available in the GNU Scientific Library [16]. In our calculations we choose to implement F_{GB2} through numerical integration using the appropriate routines from the GNU Scientific Library well aware of the resulting performance degradation.

In addition we checked the more recent 5-parameter beta distribution given in [15] that introduces a fifth parameter $0 \leq c \leq 1$, coinciding with the GB2-distribution for $c = 1$ and the GB1-distribution for $c = 0$. Parameter fitting did not result in a significant improvement and lead to a value of $c = 0.9999$ indicating that the GB2-distribution is the best choice within the 5-parameter beta distribution.

4. Technical implementation

Distribution fitting is a common task and many tools and libraries are readily available. It is – seemingly – trivial to implement the PDF and CDF of a given distribution. However, the extent of the heavy tail challenges the numerical evaluation of rather common functions (see section 4.1). For parameter fitting we use an optimization library implementing a multi-dimensional simplex-algorithm (section 4.2) and risk remaining stuck in a local optimum.

4.1 Function evaluation

It is straightforward to compute both distributions numerically: Besides elementary functions, the dPlN-distribution depends only on the error function (which is available in the GNU C math library). The GB2-distribution in turn depends on the Beta function $\mathrm{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ that can either be computed numerically using the GNU Scientific Library [16] or with the Γ-function available in the C math library.

In practice the straightforward implementation of the formulas is problematic for numerical calculations. Double precision floating point arithmetic [17] does not favor e.g. the multiplication of a term very close to zero (e.g. taking the Gaussian in equation 2) with another term asymptotically approaching 1. Lacking a numerically stable implementation of f_{dPlN} and F_{dPlN} we notice that the trivial implementation of the functions erroneously yields zero for one of the addends in the addition apparently without impacting the result. Looking at f_{GB2} the GSL Scientific Library provides a numerical implementation of the Beta function. However, in our case the computation ceases to return non-zero values for small function values of f_{GB2}. As a consequence, f_{GB2} was zero for $x > 26000$ ms using the built-in double precision floating point arithmetic.

Table 1: Results of the parameter fitting.

model	parameters
GB2	$\alpha = 166.413$, $\beta = 344.026$
$\Delta = 13.979$	$p = 0.0115766$, $q = 0.0089585$
dPlN	$\alpha = 1.51172$, $\beta = 1.85461$
$\Delta = 13.390$	$\nu = 5.85733$, $\tau = 0.150105$

As a workaround we implemented both distributions using the multi-precision software floating point library MPFR [18]. For the dPlN-distribution we dynamically increase the precision until all addends in the sum are non-zero – in rare cases leading to several hundred bit floating point precision calculations. In contrast a moderate increase in precision (96 bit floating point) is sufficient for the computation of the GB2-distribution. Of course, the consequence is a large performance degradation by more than a factor of 1 000.

4.2 Parameter fitting

To find the best parametrization we fit the PDFs f_{dPlN} and f_{GB2} to the empirical PDF (scaled to the proportion of successful connection terminations and using a 20ms binning to hide short-term network effects). To emphasize the heavy-tail we choose to calculate the weighted difference (rather than its square) between the empirical and mathematical distribution (see equation 5):

$$\Delta(\alpha, \beta, \nu, \tau) = \int_{x_{min}}^{x_{max}} \mathrm{d}x x \Big| f_{\text{empirical}}(x - x_{min}) - f_{\text{dPlN}}(x - x_{min}, \alpha, \beta, \nu, \tau) \Big|. \quad (5)$$

The fitting is limited to the region $x \in [x_{min}, x_{max}]$ and the integral is replaced by a sum over the points in time as given by the empirical data. We choose the first data point at the time difference x_{min} where the number

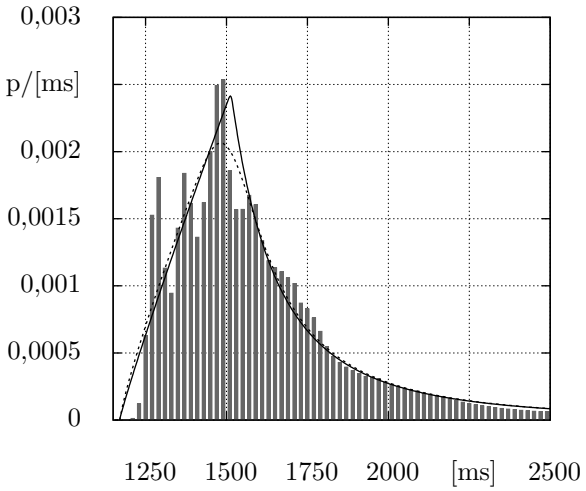

Figure 3: Histogram of the empirical probability density function with 20ms binning and the best parameter fits for the dPlN- and GB2-distribution (dashed and solid line respectively).

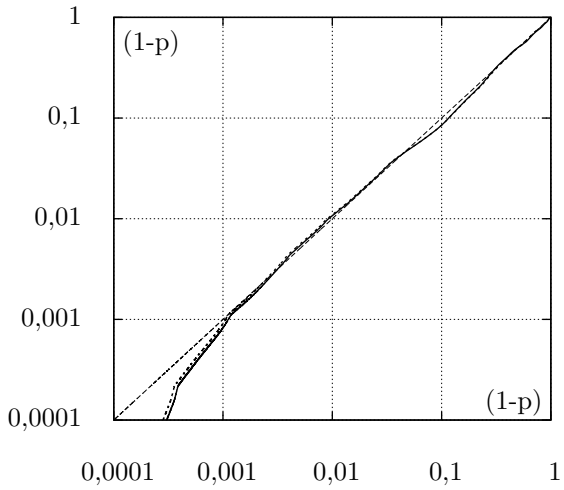

Figure 4: pp-plot comparing the empirical data and the best parameter fits for the dPlN- and GB2-distribution (dashed and solid line respectively) by plotting the complementary CDFs (Φ^c_{model},$\Phi^c_{\text{empirical}}$). A perfect fit would produce the diagonal line (shown to guide the eye).

of events in the 20ms bin exceeds 25 for the first time. In the example this leads to $x_{min} = 1170$ms. The upper limit is set to $x_{max} = 20\,000$ms. Note that the weight function in equation 5 is x (i.e. *not* $x - x_{min}$). The same setup is used to fit f_{GB2}.

For the parameter fitting we take the numerical implementation (either with hardware or software floating point arithmetic) and subject it to the minimization algorithm provided by the GSL Scientific Library (using a multi-dimensional simplex-algorithm with randomized initialization). To avoid getting stuck in a local minimum we restart the minimization repeatedly with slightly changed initial conditions until no further progress is seen within a fixed number of restarts.

The results are given in table 1: For practical purposes the fitness is indistinguishable for the GB2- and dPlN-distribution. Both distributions give a very similar asymptotic power-law behavior with a leading exponent of $-\alpha q - 1 = -2.491$ for f_{GB2} and $-\alpha - 1 = -2.512$ for f_{dPlN}.

The data has not a single, smooth peak probability: Looking at the PDF the histogram in figure 3 we see the rapid onset followed by at least three pronounced peaks before the onset of the heavy-tail. The histogram uses 20ms binning to suppress the dynamics due to the mobile network and the probability density is normalized to 1ms binning (the maximum time resolution of the log data). The existence of several peaks is no surprise since the data set encompasses four different OBU types operating in three national mobile networks and it is well known that the latency strongly depends e.g. on the type of mobile unit used [19].

Splitting the data into distinct groups as suggested by the OBU type and mobile network operator does not improve the quality of the fit but would introduce many more fit parameters – clearly deteriorating the quality of the model [20].

5. Connection termination latency

Looking at the CDF in figure 2 we note that the probability remains close to zero up to approx. 1200ms rising quickly thereafter, a picture consistent with the data observed in [5]. Comparing the latency with the data reported in [5] we deduce that four TCP packets are exchanged in our example. However, without access to the real-world system or network monitoring we cannot confirm this directly.

Both the CDF and the PDF (figures 2 and 3 respectively) show a good match between the empirical data and the dPlN- and GB2-distributions. The visibility of the remaining differences depends (strongly) on the visualization used: Whereas the CDF tends to hide differences, they are obvious in the histogram.

Looking at the tail of the distribution we choose the pp-plot of the complementary CDFs (see figure 4), i.e. plotting

$$\left(\Phi^c_{\text{model}}(x), \Phi^c_{\text{empirical}}(x)\right)$$

for $x \in [1170\text{ms}, 70000\text{ms}]$. An ideal fit would produce a diagonal line (shown in figure 4 to guide the eye) where the coordinate (1,1) corresponds to the start of the

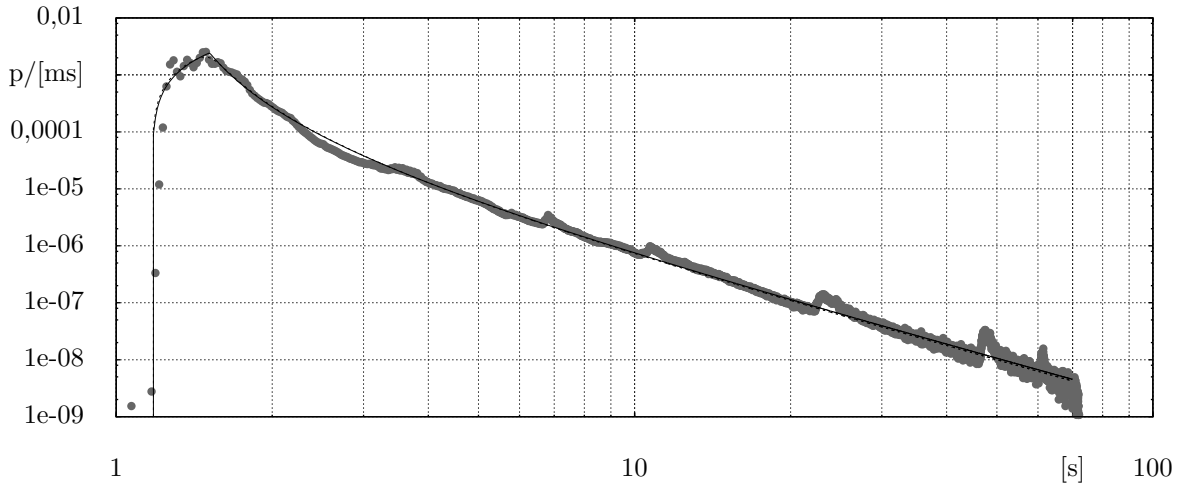

Figure 5: Probability density function (normalized to 1ms time resolution) for the TCP connection termination latency in 2G mobile networks. Empirical data (gray dots using 20ms binning) and fitted dPlN- and GB2-distributions (dashed and solid lines, fitted in the interval [1170 ms; 20 000ms]).

CDF (i.e. $\Phi = 0$) moving to the lower left corner with increasing CDF. Both the dPlN- and GB2-distribution (dashed and solid line respectively) leave the body of the distribution with moderate deviations in the vicinity of the cumulated probability of $p = 0.9$ to follow the power-law tail closely up to $p = 0.999$. Beyond that cumulated probability the fitted distributions no longer trace the empirical data.

One advantage of the GB2-distribution is that it encompasses many well-known distributions as special cases – that are automatically ruled out with the parameters fitted in our case. For some distributions we have cross-validated this by independently fitting the gamma-, lognormal-, Weibull- and Fisk-distribution. In addition we checked the Pareto- and Dagum-distributions, all to no avail.

A log-log-plot of the PDF is a simple way to visually verify the power-law behavior. Figure 5 gives the probability density function (again normalized to time intervals of 1ms, the maximum time resolution of the log data). The empirical data is shown as grey dots using a 20ms binning across the whole time range. As the latency increases the number of events per bin decreases, easily noticeable by the increasing 'noisiness' of the data. Even with more than 300 million events in the data set, at a latency of 70s the number of events per 20ms bin is getting so small that it fluctuates between approx. 10 and 30 events. 70s after the peak in probability a connection timeout is triggered and reduces the event rate by almost a factor of 100 (not shown).

Both fitted distributions describe the tail equally well: f_{GB2} with a slope of -2.491 (dashed line) and f_{dPlN} with a slightly steeper slope of -2.512 (solid line in figure 5). While the overall trend is well reproduced, several notable features are present in the empirical data. Close to 7s the first pronounced peak is visible (and present in any combination of OBU type and mobile network operator), we interpret it as the successful retransmission of one TCP packet. The following three peaks at 11s, 23s and 47s are an artifact of the older OBU types and match the spacing of an exponential-back-off algorithm potentially used in the calculation of retransmission timeouts. It remains to be seen whether sufficient events can be collected to search for the next peak close to 100s.

6. Summary

We have gathered a large data set on the connection termination latency in a 2G mobile network. Fitting the data with well-known statistical distributions we find that the dPlN- and GB2-distributions fit the data equally well, both requiring four parameters. Fitting the GB2-distribution automatically excludes many other statistical distributions as candidates since they are special cases of the GB2-distribution.

The empirical data shows that the power-law tail extends to time scales of almost 100 times the average connection termination latency before a connection timeout reduces the event rate drastically. The fitted distributions suggest that the latency distribution would otherwise extend even further.

In practice the fitted distributions apply to network simulation models and allow the simple integration of network-wide TCP latencies for 2G mobile networks, e.g. in our model of the German automatic toll system.

References

[1] A. A. Toda, "The double power law in income distribution: Explanations and evidence," *Journal of Economic Behavior & Organization*, vol. 84, no. 1, pp. 364–381, Sep. 2012. DOI: 10.1016/j.jebo.2012.04.012.

[2] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, "Mobile call graphs: Beyond power-law and lognormal distributions," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08, Las Vegas, Nevada, USA: ACM, 2008, pp. 596–604, ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401963.

[3] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009. DOI: 10.1137/070710111.

[4] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet Mathematics*, vol. 1, no. 2, pp. 226–251, Jan. 2004. DOI: 10.1080/15427951.2004.10129088.

[5] P. Benko, G. Malicsko, and A. Veres, "A large-scale, passive analysis of end-to-end TCP performance over GPRS," in *IEEE INFOCOM 2004*, vol. 3, 2004, pp. 1882–1892. DOI: 10.1109/INFCOM.2004.1354598.

[6] J.-O. Engler and S. Baumgärtner, "Model choice and size distribution: A Bayequentist approach," University of Lüneburg, working paper 265, 2013. [Online]. Available: www.leuphana.de/institute/ivwl/publikationen/working-papers.html (visited on 04/19/2017).

[7] W. J. Reed and M. Jorgensen, "The double Pareto-lognormal distribution – A new parametric model for size distributions," *Communications in Statistics – Theory and Methods*, vol. 33, no. 8, pp. 1733–1753, Sep. 2004. DOI: 10.1081/STA-120037438. [Online]. Available: https://www.math.uvic.ca/faculty/reed/dPlN.3.pdf (visited on 04/25/2017).

[8] W. J. Reed, "The normal-Laplace distribution and its relatives," in *Advances in Distribution Theory, Order Statistics, and Inference*, N. Balakrishnan, E. Castillo, and J. M. Sarabia, Eds. Boston, MA, USA: Birkhäuser, 2006, pp. 61–74. DOI: 10.1007/0-8176-4487-3_4.

[9] J. H. Graham, D. T. Robb, and A. R. Poe, "Random phenotypic variation of yeast (saccharomyces cerevisiae) single-gene knockouts fits a double Pareto-lognormal distribution," *PLOS ONE*, vol. 7, no. 11, pp. 1–6, Nov. 2012. DOI: 10.1371/journal.pone.0048964.

[10] C. C. Zhang, "The double Pareto-lognormal distribution and its applications in actuarial science and finance," Master's thesis, Université de Montréal, 2015.

[11] J. B. McDonald, "Some generalized functions for the size distribution of income," *Econometrica*, vol. 52, no. 3, pp. 647–663, May 1984, ISSN: 00129682. DOI: 10.2307/1913469.

[12] Y. Ye, "Properties of weighted generalized beta distribution of the second kind," Master's thesis, Georgia Southern University, 2012. eprint: http://digitalcommons.georgiasouthern.edu/etd/1017.

[13] Y. Ye, B. O. Oluyede, and M. Pararai, "Weighted generalized beta distribution of the second kind and related distributions," *Journal of Statistical and Econometric Methods*, vol. 1, no. 1, pp. 13–31, Feb. 2012, ISSN: 2241-0376. DOI: 10419/58014.

[14] J. Otieno, "From the classical beta distribution to generalized beta distributions," Master's thesis, University of Nairobi, 2013. DOI: 11295/52393.

[15] J. B. McDonald and Y. J. Xu, "A generalization of the beta distribution with applications," *Journal of Econometrics*, vol. 66, no. 1, pp. 133–152, Oct. 1995, ISSN: 0304-4076. DOI: 10.1016/0304-4076(94)01612-4.

[16] B. Gough, *GNU Scientific Library Reference Manual*, 3rd. Network Theory Ltd., 2009, ISBN: 9780954612078. [Online]. Available: https://www.gnu.org/s/gsl/manual/gsl-ref.pdf.

[17] D. Goldberg, "What every computer scientist should know about floating-point arithmetic," *ACM Comput. Surv.*, vol. 23, no. 1, pp. 5–48, Mar. 1991, ISSN: 0360-0300. DOI: 10.1145/103162.103163.

[18] R. Brent and P. Zimmermann, *Modern Computer Arithmetic*. New York, NY, USA: Cambridge University Press, 2010, ISBN: 9780521194693.

[19] P. Romirer-Maierhofer, F. Ricciato, A. D'Alconzo, R. Franzan, and W. Karner, "Network-wide measurements of TCP RTT in 3G," in *Traffic Monitoring and Analysis: First International Workshop, TMA 2009, Aachen, Germany, May 11, 2009. Proceedings*, M. Papadopouli, P. Owezarski, and A. Pras, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 17–25, ISBN: 978-3-642-01645-5. DOI: 10.1007/978-3-642-01645-5_3.

[20] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974, ISSN: 0018-9286. DOI: 10.1109/TAC.1974.1100705.