A Novel Probability Distribution for Modeling Internet Traffic and its Parameter Estimation

Edward Chlebus and Gautam Divgi
Network Modeling and Teletraffic Analysis Lab
Department of Computer Science
Illinois Institute of Technology
Chicago, IL - 60616
Email: {chlebus, divggau}@iit.edu

Abstract—A new probability distribution following from Lavalette’s law has been proposed for modeling wireless Internet access session statistics collected in a public nationwide Wi-Fi network. We have derived maximum likelihood estimators for this distribution and found them for three data sets of session duration and traffic volume measurements. Goodness of all fits is nearly perfect. The introduced model outperforms the Pareto, truncated Pareto and modified truncated Pareto distributions. Its flexibility makes it a good candidate for describing both long-tailed and non-long-tailed data.

I. INTRODUCTION

As the wireless Internet market is booming and the number of its customers keeps growing at a very high rate, there is an urgent need for developing efficient models, methods and tools for an analysis of exploding wireless traffic.

The long-tailed Pareto distribution [7,15] has been commonly used for characterizing many wireline Internet related network metrics [9,11]. Adopting this distribution for a closely related wireless area seems like an obvious and natural way of extending the scope of application of this well-known model.

Such attempts have been recently made in [4-6], which report using the Pareto distribution for modeling the session length. Balachandran et al. [4] didn’t provide many details of the conducted analysis but Balazinska and Castro [5] and Blinn et al. [6] rounded up short sessions to a 5-10 min. interval or omitted them from the analyzed data set, respectively, due to a 5-minute measurement sampling interval. As we will see, modeling short sessions constitutes indeed a major difficulty but neglecting them oversimplifies the problem under study.

Papadopouli et al. tried to replace the Pareto with the more general 4-parameter BiPareto distribution in [13]. In our previous paper [8] based on field data analysis we have also concluded that the Pareto distribution is not suitable for modeling session duration and traffic volume. Downey [11] has recently found that the evidence for long tails isn’t convincing even in data sets that have already been reported as long-tailed. In many cases the lognormal model describes better than Pareto the tail of observed empirical distributions.

It looks like the Pareto model, although widely accepted, needs to be reexamined. The objective of this paper is to propose a new probability distribution for modeling Internet traffic, estimate its parameters and validate against measurement data.

II. MEASUREMENT DATA

The traffic measurement statistics used for this analysis were provided by Azure Wireless [3] one of the biggest Wi-Fi hotspot network operators in Australia. We have selected for our study customers of the 30-day prepaid accounts. They are the most active in terms of the number of initiated sessions and, unlike the owners of e.g. 1-day or 1-hour accounts, are not psychologically forced to make the most of their limited allowed Internet access time, which may cause a behavior atypical for an average user. Traffic statistics were collected over 5 months from all network access points located all over Australia (see [3] for details of hotspot locations).

A wireless Internet access session is under our investigation. By a session we mean the total time from logging in to logging out, over which a user has access to the network. For every session its duration and up- and downloaded traffic volumes are given in three separate data sets extracted from the log, from now on referred to as A, B and C, respectively. They have already been investigated in our previous papers [8,10]. There are $n=2625$ samples in each of the analyzed data sets.

Their basic characteristics are given in Tab. 1 of [8].

The measurement resolution was 1 s and 1 byte for a session length and traffic volume, respectively. Note that the duration of a session was limited to 5 hours by the network operator. Our goal was to investigate sessions intentionally terminated by users so all sessions whose length is equal to exactly 18000 seconds, i.e. forced to termination by the 5-hour timeout, have been disregarded in the analysis.

III. FITTING THE PARETO AND TRUNCATED PARETO DISTRIBUTIONS TO THE DATA

The long-tailed Pareto distribution with the probability density function (PDF) and cumulative distribution function (CDF) for $0 < x \leq \infty$ and $\alpha > 0$ given by [7,15]

$$f(x) = \alpha \cdot x^{-(\alpha+1)}$$

and

$$F(x) = 1 - \left( \frac{k}{x} \right)^{\alpha}$$

1930-529X/07/$25.00 © 2007 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.
respectively, is frequently used for modeling many Internet features. A complementary cumulative distribution function (CCDF) \( F_n(x) = 1 - F(x) \) of a Pareto distributed random variable \( X \) results in a straight line on a log-log scale. Unfortunately this is not the case for any of the investigated data sets A, B or C. The optimal fits determined for the examined \( n \) samples of \( X \) by the maximum likelihood estimators \([7,15]\)

\[
k = \min_i x_i \quad (2a)
\]

\[
\alpha = \frac{n - 1}{\sum_{i=1}^{n} \ln \frac{x_i}{k}} \quad (2b)
\]

are depicted in Figs. 1-3 (see Tab. I for the numerical values of \( k \) and \( \alpha \)). As we can see they are extremely poor.

To eliminate subjectivity of visual plot examination, goodness of the Pareto fit has been evaluated by applying the Kolmogorov-Smirnov (KS) supremum statistic [2]

\[
D = \sup_x |F_n(x) - F(x)| \quad (3)
\]

where \( F_n(x) \) and \( F(x) \) denote the empirical and fitted analytical CDFs, respectively. The numerical values of \( D \) are given in Tab. I. We support our conclusion that the investigated data sets don’t follow the Pareto distribution. Note that a smaller value of \( D \) indicates a better fit.

It is obvious that the Pareto model doesn’t work. One of possible reasons may be the 5-hour timeout imposed on the session length by the network operator. If so, the upper truncated Pareto distribution whose PDF and CDF for \( 0 < k < x < m < \infty \), \( k < m \) and \( \alpha > 0 \) are given by [1]

\[
f(x) = \frac{\alpha x^{-\alpha - 1}}{1 - \left( \frac{k}{m} \right)^\alpha} \quad (4a)
\]

and

\[
F(x) = \frac{1 - \left( \frac{k}{x} \right)^\alpha}{1 - \left( \frac{k}{m} \right)^\alpha} \quad (4b)
\]

respectively, may be a reasonable alternative to the unsuccessful Pareto distribution. Unfortunately, the maximum likelihood fits with the estimators [1]

\[
k = x_{\min} = \min(x_1, x_2, \ldots, x_n) \quad (5a)
\]

\[
m = x_{\max} = \max(x_1, x_2, \ldots, x_n) \quad (5b)
\]

and \( \alpha \) solving the equation

\[
\frac{n}{\alpha} \left( \frac{x_{\min}}{x_{\max}} \right)^\alpha \ln \frac{x_{\min}}{x_{\max}} - \sum_{i=1}^{n} (\ln x_i - \ln x_{\min}) = 0 \quad (5c)
\]

are still not satisfactory (see Figs. 1-3 and Tab. 1).

IV. MODIFYING THE TRUNCATED PARETO DISTRIBUTION

Consider \( n \) samples of the random variable \( X \) sorted in an increasing order

\[
x_{s} \leq x_{s+1} \leq \ldots \leq x_{n} \leq x_{1} \quad (6)
\]

The index \( r \) in the above formula denotes the rank of the sample \( x_r \) with \( x_1 = x_{\max} \) and \( x_n = x_{\min} \). Figs. 7-8 present \( x_r \) as a function of its rank \( r \) on a linear-log scale. They show a clear linear trend \( x_r = g(r) \) for most of the range \( r_1 \leq r \leq r_u \), where \( r_1 \) and \( r_u \) denote the lower and upper bound of the linear region, respectively. We have proved in our paper [8] that the samples

\[
x_{s} \leq x_{s+1} \leq \ldots \leq x_{s+n+2} \leq x_{s+n+1} \leq x_{s+n} \quad (7)
\]

come from the probability distribution which is a special case of the truncated Pareto distribution (4) with \( \alpha = 0 \). Its CDF has the form \([8]\)

\[
F(x) = \frac{\ln x}{x} = \frac{\ln x}{x} - 1 - \left( \frac{k}{x} \right)^\alpha \quad (8a)
\]

\[
F(\ln m) = \frac{\ln m}{\ln k} - 1 \quad (8b)
\]

To apply this distribution to all the data samples (6) we modify the values of \( k \) and \( m \) by extending the linear trend \( g(r) \) for the entire domain of \( r \), i.e. \( k = g(r = n) \) and \( m = g(r = 1) \). The CDF (8) of the modified truncated Pareto distribution finally becomes

\[
F(x) = \frac{\ln x}{x} = \frac{\ln g(n)}{\ln g(1)} \quad (9)
\]

Details of \( g(r) \) are discussed in [8].

The CDF (9) fits very well to the body of the empirical distribution but, as expected, its performance deteriorates for both lower and upper tails that don’t follow the linear trend \( g(r) \) (cf. Figs. 7-8). The KS statistics of (9) given in Tab. 1 are superior to the previously tested distributions (1) and (4) but still not good enough. We need a better model.

V. RELATING DATA SAMPLES WITH THEIR RANKS

Extending the linear trend \( x_r = g(r) \) discussed in the previous section for the nonlinear regions provides only a partial solution to the problem. The resulting probability distribution (9) performs best as yet, but due to significant deviations from \( g(r) \) for both low and high ranks \( r \), it is still inappropriate for describing the analyzed field data.

In order to accurately relate \( x_r \) with \( r \) we adopt Lavalette’s law [12] (see also [14])

\[
x_r = c \left( \frac{n + 1 - r}{r} \right)^\beta \quad (10)
\]

originally proposed to express \( n \) ordered values of the journal impact factor \( x_r \) as a function of their ranks \( r \). \( c \) and \( \beta \) are parameters to be determined. The formula (10) describes very well a characteristic semi-logarithmic S-shape curve [14] resembling the empirical ones shown in Fig. 8. It looks like trying to apply Lavalette’s law to our data sets may be worthwhile. In order to increase fitting capability of (10) we introduce an additional parameter \( k \) in the denominator. This gives the formula
\[ x_r = c \left( \frac{n+1-r}{k+r} \right)^\beta \]  
(11)

Taking the logarithm of both sides of (11) we get

\[ \log x_r = \beta \log \frac{n+1-r}{k+r} + \log c \]  
(12)

so \( x_r \) is a linear function of

\[ z = \frac{n+1-r}{k+r} \]  
(13)

if both \( x_r \) and \( z \) are presented on a logarithmic scale. Given \( k \) we can use linear regression \([2]\), successfully applied in our earlier papers \([8,9]\), to find a straight line that fits best to a log transformed data set. Since \( k \) is unknown we search for optimal values of the parameters \( \beta \) and \( c \) by maximizing the coefficient of determination \( R^2 \)\([2]\) with respect to \( k \). The higher value of \( R^2 \), the better goodness of a linear fit. This optimization has been run for integer values of \( k \) to speed up the process and is illustrated in Fig. 9. The resulting optimal values of all parameters are listed in Tab. II. We can see from it that introducing the additional parameter \( k \) in the original Lavalette’s law has been justified. All \( R^2 \) values are higher for (11) than for (10). Note that quality of all fits is excellent (\( R^2 = 1 \) indicates a perfect one). They are depicted in Figs. 7-8.

VI. INTRODUCING A NOVEL PROBABILITY DISTRIBUTION FOR MODELING INTERNET TRAFFIC

Following (6) the empirical cumulative distribution function of the random variable \( X \) is given by

\[ F_c(x_r) = P(X \leq x_r) = \frac{n-r+1}{n} \]  
(14)

From (11) we get

\[ r = \frac{n+1-k}{\left( \frac{x_r}{c} \right)^\beta + 1} \]  
(15)

Substituting this for \( r \) in (14) yields

\[ F_c(x_r) = 1 + \frac{1}{n} - \frac{k}{n} \left( \frac{x_r}{c} \right)^\beta \]  
(16)

The CDF for a continuous domain of \( x \) is obtained as a limit case for \( n \to \infty \)

\[ F(x) = \lim_{n \to \infty} F_c(x) = 1 - \frac{k}{n} \left( \frac{x}{c} \right)^\beta \left( 1 + d \right) = 1 + d \left( \frac{x}{c} \right)^\beta \]  
(17a)

where \( d = \frac{k}{n} \). From \( F(x_{\text{max}}) = 1 \) one can easily find

\[ d = \left( \frac{c}{x_{\text{max}}} \right)^\beta \]  
(17b)

Differentiating both sides of (17a) with respect to \( x \) we get PDF of \( X \)

\[ f(x) = \frac{1}{1 + d \left( \frac{x}{c} \right)^\beta} \]  
(17c)

VII. MAXIMUM LIKELIHOOD ESTIMATORS FOR THE NOVEL DISTRIBUTION

The likelihood function \([2]\) for the proposed probability distribution (17)

\[ l(x_1, x_2, ..., x_n; c, \beta, d, \theta) = f(x_1)f(x_2) \cdots f(x_n) \]

\[ = \left( 1 + d \right) \frac{1}{\left( \frac{x}{c} \right)^\beta} \]  
(18)

immediately implies the following logarithmic likelihood function

\[ L(x_1, x_2, ..., x_n; c, \beta, d, \theta) = \ln(l(x_1, x_2, ..., x_n; c, \beta, d, \theta)) \]

\[ = n \ln(1 + d) - n \ln \beta - n \ln c + \sum_{i=1}^{n} \ln x_i - 2 \sum_{i=1}^{n} \ln \left( \frac{x_i}{c} \right)^\beta + 1 \]  
(19)

Rewriting (19) with (17b) eliminates \( d \) and reduces the number of estimated parameters from three to two, i.e. only \( c \) and \( \beta \)

\[ L(x_1, x_2, ..., x_n; c, \beta) = n \ln \left( 1 + \left( \frac{x_{\text{max}}}{c} \right)^\beta \right) - n \ln \beta - n \ln c + \]

\[ + \frac{1}{\beta} \sum_{i=1}^{n} \ln x_i - 2 \sum_{i=1}^{n} \ln \left( \frac{x_i}{c} \right)^\beta + 1 \]  
(20)

To find the optimal values of \( c \) and \( \beta \) maximizing the likelihood function (20) we set the partial derivatives equal to zero

\[ \frac{\partial L}{\partial \beta} = \frac{1}{\beta^2} \frac{n}{\left( \frac{x_{\text{max}}}{c} \right)^\beta} = 0 \]  
(21a)

\[ \frac{\partial L}{\partial c} = \frac{1}{\beta^2} \frac{n}{\left( \frac{x_{\text{max}}}{c} \right)^\beta} \ln c - \frac{n}{\left( \frac{x_{\text{max}}}{c} \right)^\beta} \beta = 0 \]  
(21b)

The introduced functions \( h_c(c, \beta) \) and \( h_\beta(c, \beta) \) denote the second factor of the partial derivatives (21a) and (21b), respectively.

1930-529X/07/$25.00 © 2007 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.
Since for a fixed value $\beta = \beta_f$, $h_c(c, \beta_f)$ is a monotonically decreasing function of $c$ (see Fig. 10), the equation $h_c(c, \beta_f) = 0$ can be solved for $c$ by the following iterative bisecting search algorithm:

**Algorithm 1**

Step 1. Set the initial lower $c_l$ and upper $c_u$ bounds for $c$.

Step 2. For a fixed $\beta_f$ and $c = \frac{c_l + c_u}{2}$ calculate the numerical value of $h_c(c, \beta_f)$. If $|h_c(c, \beta_f)| < \varepsilon$ then stop, else go to Step 3.

Step 3. If $h_c(c, \beta_f) > 0$ update the lower bound $c_l = c$ and go to Step 2, else update the upper bound $c_u = c$ and go to Step 2.

The Algorithm 1 uniquely determines $c_\beta$ for a given $\beta$ so $h_{c_\beta}(c_\beta, \beta)$ becomes a function of the variable $\beta$ only. We can see from Fig. 11 that $h_{c_\beta}(c_\beta, \beta)$ also decreases monotonically as $h_c(c, \beta_f)$ which suggests solving the equation $h_{c_\beta}(c_\beta, \beta) = 0$ by an algorithm similar to the one proposed above:

**Algorithm 2**

Step 1. Set the initial lower $\beta_l$ and upper $\beta_u$ bounds for $\beta$.

Step 2. For $\beta = \frac{\beta_l + \beta_u}{2}$ find $c_\beta$ by applying Algorithm 1.

Step 3. Calculate the numerical value of $h_{c_\beta}(c_\beta, \beta)$. If $|h_{c_\beta}(c_\beta, \beta)| < \varepsilon$ then stop, else go to Step 4.

Step 4. If $h_{c_\beta}(c_\beta, \beta) > 0$ update the lower bound $\beta_l = \beta$ and go to Step 2, else update the upper bound $\beta_u = \beta$ and go to Step 2.

The maximum likelihood estimators $c$, $\beta$ and $d$ (cf. Eq. 17b) have been determined with the numerical accuracy $\varepsilon = 10^{-12}$. Their optimal values and goodness of the resulting fit are compared with the previously evaluated models in Tab. 1. Figs. 4-6 illustrate the corresponding CDFs.

**VIII. CONCLUSIONS**

The new probability distribution (17) proposed in this paper provides an excellent fit to the wireless Internet measurement data and significantly outperforms all other investigated models. The resulting Kolmogorov-Smirnov supremum statistics are very low and the fitted CDFs nearly perfect. They overlap the corresponding empirical curves for all the examined data sets. The model performance is excellent for both the session traffic volume revealing the S-shape (cf. Fig. 8) typical for the original application of Lavalette’s law and the session duration, that unlike the traffic volume, doesn’t display significant upward deviation from the linear trend for low ranks (cf. Fig. 7).

Note that the introduced probability distribution offers great flexibility of upper tail modeling (see Fig. 12). If $d = 0$ the distribution is clearly non-long-tailed but for the case $d = 0$ and large $x$, linear behavior of the plot on the log-log scale

$$\lim_{x \to \infty} \frac{\log F_c(x)}{\log x} = -\frac{1}{\beta}$$

indicates a long tail. It looks like our model can be tuned to both long-tailed and non-long-tailed data. This is a very desirable feature since the widely applied Pareto distribution fails to model many Internet variables [11].

Estimating the parameters $c$ and $\beta$ of the model wasn’t trivial but the effort was definitely worthwhile. With respect to the results of this paper, the new probability distribution looks very promising. Investigating the range of its application will be a subject of our further study.

**REFERENCES**


TABLE I. COMPARISON OF OPTIMAL FITS PROVIDED BY EXAMINED PROBABILITY DISTRIBUTIONS

<table>
<thead>
<tr>
<th>Data set</th>
<th>Pareto (1)</th>
<th>Truncated Pareto (4)</th>
<th>Modified truncated Pareto (9)</th>
<th>Proposed distribution (17)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A – session duration</td>
<td>k = 1.00</td>
<td>m = 17998.00</td>
<td>k = 291.59</td>
<td>c = 2464.6102</td>
</tr>
<tr>
<td></td>
<td>α = 0.1336</td>
<td>D = 0.1670</td>
<td>m = 13662.66</td>
<td>d = 0.1127</td>
</tr>
<tr>
<td></td>
<td>β = -0.3962</td>
<td></td>
<td>D = 0.0930</td>
<td>β = 0.9108</td>
</tr>
<tr>
<td>B – uploaded traffic volume</td>
<td>k = 66.00</td>
<td>m = 310136246.00</td>
<td>k = 45958.99</td>
<td>c = 444095.8604</td>
</tr>
<tr>
<td></td>
<td>α = 0.1145</td>
<td>D = 0.2820</td>
<td>m = 4323325.40</td>
<td>d = 0.0023</td>
</tr>
<tr>
<td></td>
<td>β = -0.0538</td>
<td></td>
<td>D = 0.1064</td>
<td>β = 1.0792</td>
</tr>
<tr>
<td>C – downloaded traffic volume</td>
<td>k = 168.00</td>
<td>m = 488977830.00</td>
<td>k = 224033.99</td>
<td>c = 2523147.1658</td>
</tr>
<tr>
<td></td>
<td>α = 0.1053</td>
<td>D = 0.2618</td>
<td>m = 28762488.03</td>
<td>d = 0.0095</td>
</tr>
<tr>
<td></td>
<td>β = -0.1164</td>
<td></td>
<td>D = 0.1082</td>
<td>β = 1.1315</td>
</tr>
</tbody>
</table>

TABLE II. OPTIMAL PARAMETERS OF LAVELETTE’S LAW DETERMINED BY LINEAR REGRESSION

<table>
<thead>
<tr>
<th>Data set</th>
<th>Original Lavallette’s law (10)</th>
<th>Proposed Lavallette’s law (11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A – session duration</td>
<td>c = 1772.7138</td>
<td>k = 322</td>
</tr>
<tr>
<td></td>
<td>β = 0.7358</td>
<td>β = 0.9212</td>
</tr>
<tr>
<td></td>
<td>R² = 0.9524</td>
<td>R² = 0.9968</td>
</tr>
<tr>
<td>B – uploaded traffic volume</td>
<td>c = 407088.0182</td>
<td>k = 24</td>
</tr>
<tr>
<td></td>
<td>β = 1.0988</td>
<td>β = 1.1486</td>
</tr>
<tr>
<td></td>
<td>R² = 0.9829</td>
<td>R² = 0.9865</td>
</tr>
<tr>
<td>C – downloaded traffic volume</td>
<td>c = 2221710.2721</td>
<td>k = 115</td>
</tr>
<tr>
<td></td>
<td>β = 1.1061</td>
<td>β = 1.2521</td>
</tr>
<tr>
<td></td>
<td>R² = 0.9752</td>
<td>R² = 0.9918</td>
</tr>
</tbody>
</table>
Figure 7. Session duration vs. its rank (linear-log scale).

Figure 8. Traffic volume vs. its rank (linear-log scale).

Figure 9. Optimization of the parameter $k$ for the modified Lavalette’s law (11).

Figure 10. The function $h_c(c, \beta_f)$ (21a) for Data Set A (session duration).

Figure 11. The function $h_{\beta}(c_{\beta}, \beta)$ (21b) for Data Set A (session duration).

Figure 12. CCDF of the proposed probability distribution (17) for varying $d$; $c=100000$, $\beta = 1$. 

---

1930-529X/07/$25.00 © 2007 IEEE

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.