Experimental Analysis of Bittorrent Traffic based on Heavy-Tailed Probability Distributions

Aishwarya Gaikwad
Student, Dept. of E&TC
Pune Institute of Computer Technology,
Pune, India

Rupesh Jaiswal
Associ. Prof.,
Dept. of E&TC
Pune Institute of Computer Technology,
Pune, India

ABSTRACT
Complexity involved in measuring and analyzing the BitTorrent traffic has led to various studies in this direction. Challenges involved are related to storage, data retrieval, location of content, topological features, privacy, copyright issues along with analysis of data and modeling of traffic. Internet Traffic, earlier thought to be of Poisson, is bursty in nature. In this paper, BitTorrent traffic for applications like video is observed by means of distributions, that best represent their nature. Inter-arrival times and lengths of packets are the parameters used to plot cdf so that the best distribution is determined. This analysis can further help in exploring various fractal characteristics[1], as the alpha value obtained is crucial in determining the heavy tailed-ness, responsible for impacting network performance and creating obstacles in maintaining the desired QoS.

Keywords
P2P traffic, BitTorrent, log distributions, Inter arrival time, packet size, pcap file

1. INTRODUCTION
BitTorrent is gaining popularity at an incredible rate. It replaced the former client-server model. It's based on peer-to-peer model, which not only added complexities but call forth number of challenges in various aspects. Measuring share of traffic being that of P2P, analyzing the data would definitely help in understanding the overall internet traffic.

There are two methods for traffic measurements. Out of which passive measurement is considered in this study. Passive measurement involves silently supervising the network without any intrusion. This method is useful for simulation using traces, validation of models and identification of bottlenecks.

This study is performed in a linux environment with Ubuntu 15.10 OS for reliable operation and desired packets. Packet capture using software gives an accuracy of tens of microseconds. A widely used tool for live internet traffic is tcpdump[2]. Pcap library which comprises of Berkeley Packet Filter(BPF)[3] is related to this tool. This allows for applying stringent rules on the type of packet to be captured.

Moreover, Visual Inspections employing graphical plots like histograms (Experimental Probability Density Function),CCDF,EDF prove of great help to experts. In this study, histograms and CCDF are considered.

Further, Pareto, Weilbull and Log distributions are studied. Different probability distribution variants are tested for distribution fitting process. As a standard for goodness of fitting, the appropriate distribution type for packet length and packet inter-arrival time of BitTorrent traffic for video applications is decided.

The main goal is to comprehend the characteristics of BitTorrent traffic, and based on that, to develop P2P simulation models milieu.

2. BACKGROUND STUDY
2.1 Traffic Analysis
For the designing of computer networks and services, it is necessary to study internet traffic and develop reliable models for the same. Earlier, the Poisson model was adopted. However, the property of bursting died off over long time scales.[4 5 6]. Leland et al introduced the term long range dependence [7] in his research paper and underscored the fact that heavy tails, long-range dependent behavior, and self-similarity or fractal characteristics[1,8] affect the internet traffic and hence can prove useful in its analysis.

2.2 BitTorrent Overview
BitTorrent Protocol[9] is a peer-to-peer protocol, where work load and time consumed in accessing server for large files is reduced considerably. The content is shared in pieces simultaneously among the peers, thus enabling efficient distribution of network load.

BitTorrent swarm comprises of peers and at least one tracker. The peer request contains self-data and some stats to the tracker, which responds with the number of peers involved and their locations.

Brief mechanism of the protocol is as follows:

Suppose a file F is to be downloaded. Firstly, a torrent file ,from one of its websites is downloaded. It has meta-info of the required file, i.e it contains tracker's address and information pieces of required content. A BitTorrent client software opens this file , finds connection to the tracker and sends a list of randomly selected peers (approx.50). These peers connect to the remote peers and form a static network. However, while downloading simultaneously, uploading to smaller subsets of peers is also performed. Peers for uploading process are selected using the CHOKING or UNCHOKING mechanism[10]. This collection of active links forms active topology, which is changing frequently as compared to a more stable static network[11].

Using BitTorrent, it is possible to have fair and effective distribution of data. However, according to a study[12], download latency is six times larger for low bandwidth than high bandwidth, where the high bandwidth is three times the lower one. Free riding was common with earlier P2P systems[14]. Hence, fairness is achieved to some extent by using tit-for-tat policy. Here, download speed for any peer is

directly proportional to its contribution in sharing with other peers, i.e upload bandwidth.

However, the existing BitTorrent is facing problems like weak service availability, biased services to peers and unstable downloading performances. The model put forward in[13] reveals that the possible causes are exponential increment in the peer arrival rate and motivates for more inter-torrent collaborations.

2.3 Selection of Distributions

Distribution selection is an analytic process of scrutinizing EPDF and CCDF plots for a particular parameter. EPDF best fits for inter-departure and inter-arrival times whereas CCDF is used for assessing the message rates. Misjudging the rates, duration, and sizes of packets may adversely affect the network.

Initial method for distribution selection is a merger of visual inspection and effective use of data. Mostly hypothesis distributions like summary statistics and graphical methods is underscored in this study. Using summary stats, not only mean and variance, but some useful estimates like co-efficient of variation (cv) and skewness (v) are also obtained. An additional statistic called kurtosis can be used as a measure of peak. Visual inspection is done using graphical methods like alpha estimation [15], histogram CCDF, EDF and Hill. Lower quartiles of data is observed effectively using pdf plots whereas CCDF renders the same purpose for upper tail. Histogram is beneficial in observing high frequency data like inter-arrival times. Alpha estimation is useful in indicating the extent of self-similarity in data.

Various distributions and their applications are studied .Tests like Anderson-Darling, Shapiro-wilk are used to obtain moments, discussed above and in mathematical sections, to find the best-fit distribution.

Distribution denotes behavior of the process which plots the number of times the variable displays a specific value or range of values instead of the variable itself. In this paper, a heavy tailed distribution [23] is observed, where unlike exponential or normal distribution, the tail is heavy, as it obeys a power law.

Heavy-tailed distributions are not exponentially bounded. Earlier, Poisson graph decayed to zero. However, it is not possible to apply this for internet traffic. This is because, internet traffic datasets do not decay down to zero.

$$\bar{F}(x) = cx^n e^{-\lambda x} \qquad (i)$$

Power of n causes this distribution to decay slower than exponential. Heavy-tail implies that the larger values of x has non-negligible probability.

3. RELATED WORK

In[20], problems and issues with the existing P2P file sharing applications are classified according to the stack organization of network. Future performance questions can be answered based on the research done in the paper and compared with the existing analytical and mathematical modeling of these applications.

In[16], peers and tracker logs are used to analyze the performance globally and on session scales. They substantiated the tit-for-tat policy and confirmed the flexibleness and scalability of the policy.

Results of flow-fluid model presented by Qui and Srikant in[17] state that the amount of seeds and leechers are Gaussian random varibles.

Tracker log files based on session sizes, peer bandwidth and share ratios was analysed by Nicoll et al and observed that 80% of peers tend to download more than upload.

In[18], Karagiannis et al. identified patterns that connect source and destination IP addresses.

4. MATHEMATICAL ANALYSIS

4.1 Co-efficient of Variation

It measures the spread that explains the variability with respect to the mean. This is quite useful for exponential behavior because its coefficient of variation[22] is zero.

$$\boldsymbol{cv = \frac{\sqrt{S^2}}{\bar{X}}} \qquad (\mathbf{1})$$

$\bar{X} =$ Mean value

S= Variance

4.2 Skewness

At times, the distribution about mean is skewed or distorted with respect to the normal distribution. Skewness [22] at zero value indicates that it has symmetric distribution. Less than zero is left skewed whereas more than zero is right skewed.

$$\boldsymbol{v = \frac{1}{n(S^3)} \sum_{i=1}^{n} [x_i - \bar{X}]^3} \qquad (\mathbf{2})$$

n= sample values

4.3 Kurtosis

Flatness or peak of a distribution is measured using the kurtosis formula. Large kurtosis means the distribution is heavy-tailed and vice versa. Distributions with large kurtosis [22] tend to have a peak near the mean whereas flats peaks are observed for distributions with low kurtosis.

$$\boldsymbol{k = \frac{1}{nS^4} \sum_{i=1}^{n} [x_i - \bar{X}]^4} \qquad (\mathbf{3})$$

4.4 Pareto Distribution

The PDF of Pareto Distribution[19] is:

$$\boldsymbol{f_X(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \geq x_m \\ 0, & x < x_m. \end{cases}} \qquad (\mathbf{4})$$

The CDF for Pareto Distribution is:

$$\boldsymbol{F(x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^\alpha, & x \geq x_m \\ 0, & x < x_m. \end{cases}} \qquad (\mathbf{5})$$

$x_m =$ scale parameter

$\alpha =$ Shape parameter= tail index

4.5 Weibull Distribution

The PDF for three parameter Weibull distribution[19] is given as:

$$\boldsymbol{f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} e^{\left(-\left(\frac{x-\gamma}{\beta}\right)^\alpha\right)}} \qquad (\mathbf{6})$$

The CDF for three parameter Weibull distribution is given as:

$$F(x) = 1 - e^{\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)} \quad (7)$$

α= scale parameter ,γ= location parameter, β= shape parameter

4.6 Log-normal Distribution

The PDF for three parameter log-normal distribution[19,21] is given as:

$$f(x;\sigma,\mu,\gamma) = \frac{1}{(x-\gamma)\sigma\sqrt{2\pi}} e^{\frac{-1\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)^2}{2}} \quad (8)$$

The CDF for three parameter log-normal distribution is given as:

$$F(x;\sigma,\mu,\gamma) = \varphi\left[\frac{\ln(x-\gamma)-\mu}{\sigma}\right] \quad (9)$$

$$\sigma > 0, \qquad -\infty < \mu < \infty, \qquad 0 \le \gamma < x$$

φ = Laplace integral

5. DATA ACCUMULATION PROCESS

In linux Ubuntu 15.10 environment, with wired LAN connection, the packets are captured using Wireshark application, at an interval of ten minutes. System has 4GB RAM and intel i5 core processor. Five times the packets are captured in the complete download duration of 50 minutes. Download speed of torrent, at the time of capturing, was at an average 150kbps, whereas, simultaneous upload speed was 2 kbps. Qbittorent and Bitorrent client software are used.

Traces with pcap file format and dump file format are processed further for subsequent experiment to obtain packet inter-arrival time and packet size. The obtained data is then subjected to statistical tools of Matlab , in order to determine the best fit distribution.

6. RESULT ANALYSIS:

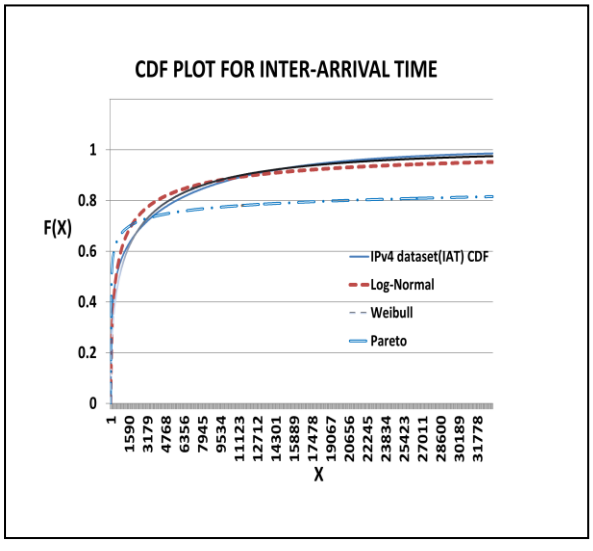

Figure 1: CDF plot for Dataset(IAT). Best fit obtained for Weibull Distribution.

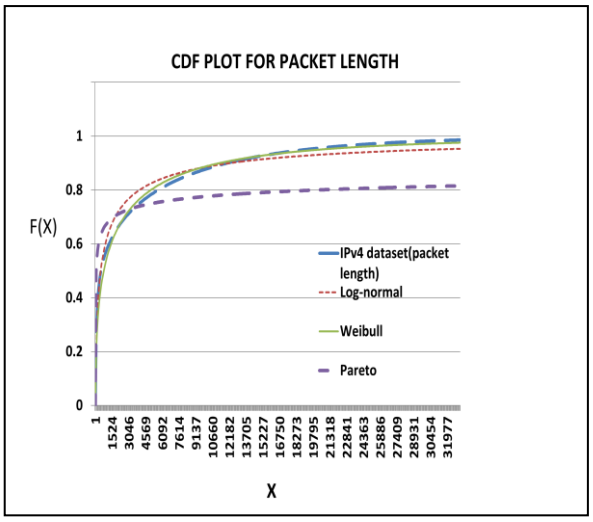

Figure 2: CDF Plot for Dataset(Packet length). Best fit obtained for Weibull Distribution.

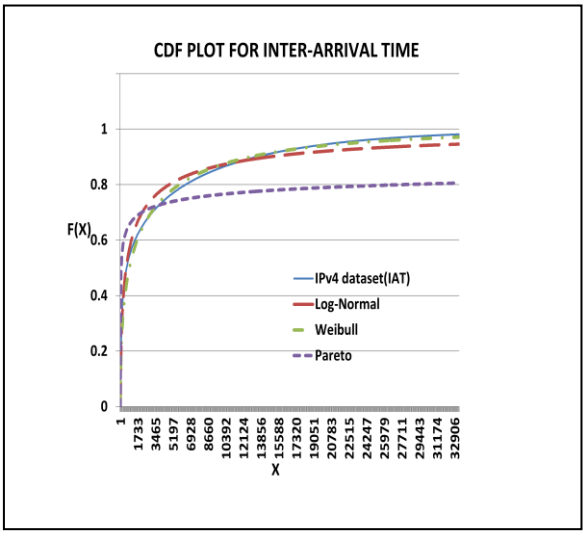

Figure 3: CDF plot for IAT Dataset. Best fit obtained for Weibull Distribution.

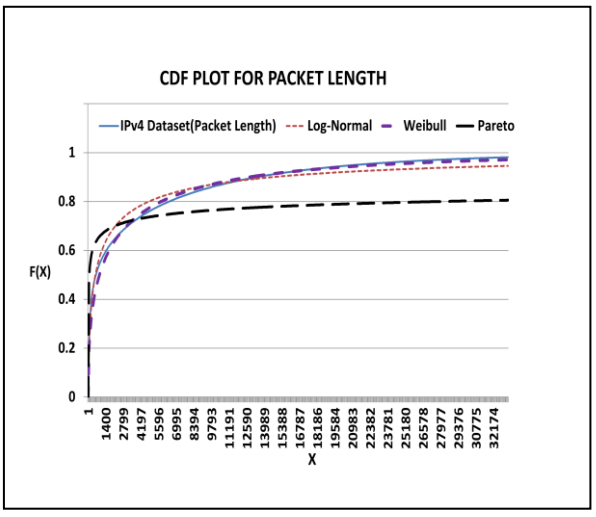

Figure 4: CDF plot for Dataset(Packet length). Best fit obtained for Weibull distribution.

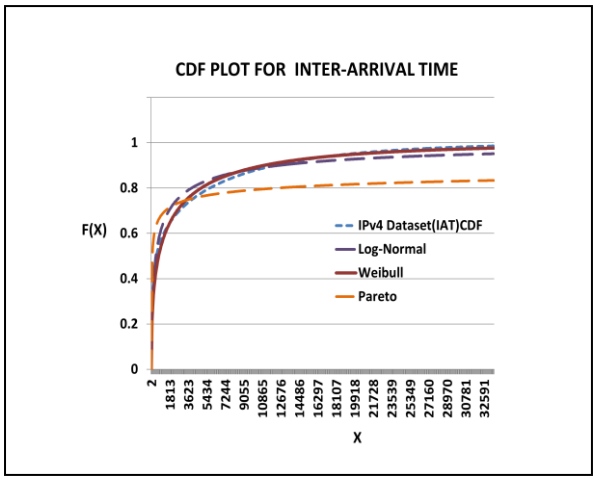

Figure 5: CDF plot for IAT dataset. Best fit obtained for Weibull distribution

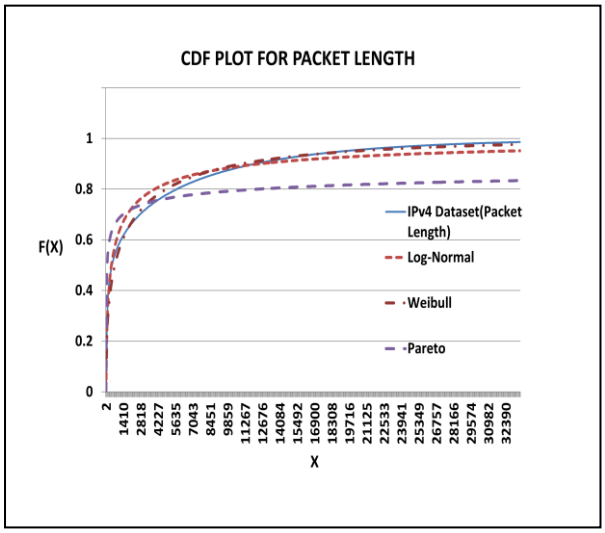

Figure 6: CDF plot for Dataset(Packet length). Best fit obtained for Weibull distribution.

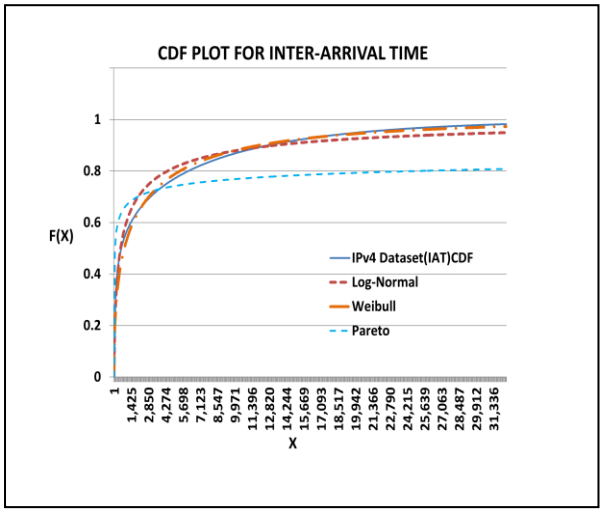

Figure 7: CDF plot for IAT dataset. Best fit obtained for Weibull distribution.

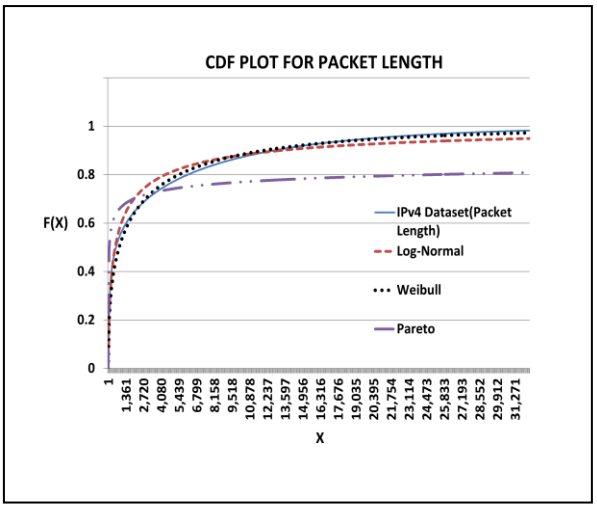

Figure 8: CDF plot for dataset(packet length). Best fit obtained for Weibull distribution.

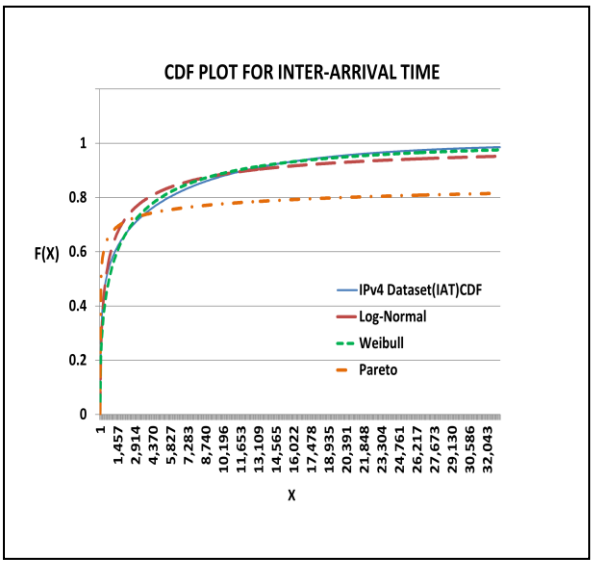

Figure 9: CDF plot for IAT dataset. Best distribution is obtained for Weibull distribution.

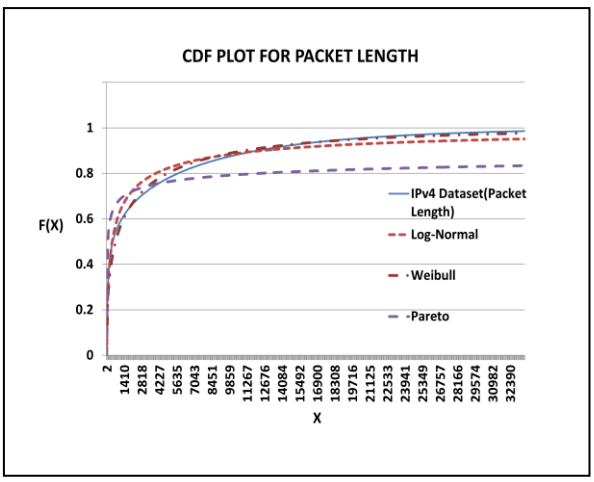

Figure 10: CDF plot for dataset(packet length). Best fit is obtained for Weibull distribution

In fig.1 and 2, pcap file of 5300 packets was used for further processing, where inter-arrival and lengths of packets were

calculated. This was further utilized to plot CDFs for each parameter.

In fig.3 and 4, with the help of another pcap file of 5250 packets, the inter-arrival and lengths of packets were evaluated so that these two parameters are used to plot CDFs in respective figures.

Again, a pcap file of 5660 packets was captured to obtain the two parameters: inter-arrival and packet lengths of packets. These were then plot as CDFs in fig. 5 and 6 respectively.

Similarly, fig. 7 and 8 are the CDF plots of inter-arrival and length of packets resp. These parameters are obtained after processing the pcap file of 5890 packets.

And finally, processing of pcap file of 5560 packets using a JAVA application, results in inter-arrival and lengths . These are then use to plot CDFs using statistical tools of MATLAB in fig 9 and 10 resp.

In all the figures, we found out that Weilbull is best fit heavy-tailed distribution for Bittorrent traffic.

8. CONCLUSION

In this paper, we examined the characteristics of BitTorrent traffic. This was performed in the Linux Ubuntu environment, with wired Internet connection. The inter-arrival time and packet length were obtained from the packets captured on Wireshark. Detailed analysis was carried out in terms of CDF.

From the outputs, we come to know that the Poisson model, which was earlier used, is unfit for the analysis of BitTorrent traffic. Instead heavy tailed-distributions like Weibull is seen to be suitable for this purpose. Many important terms related to heavy-tailed distributions like skewness, co-efficient of variation and kurtosis are studied. Alpha value, which is obtained in the analysis, measures the amount of heavy-tailedness.

This analysis can prove useful for network provider to design and manage internet traffic efficiently, in order to avoid unnecessary deterioration of QoS, caused by heavy-tailed-ness.

9. REFERENCES

[1] W.Gong, Y.Liu, V. Misra and D.Towsley. Self-Similarity and Long-range Dependence on the Internet: A Second look at the evidence, origins and Implications, Computer Networks, vol 48, No.3,pp. 377-399,2005.

[2] V.Jacobsen, C.Leres, and S.McCanne Tcpdump. http://www.tcpdump.org, August 2005.

[3] Steven McCanne and Van Jacobson. The BSD packet filter: A new architecture for user-level packet capture. In USENIX Winter, pages 259–270, 1993.

[4] M.Crovella and A.Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes, IEEE/ACM Transactions on Networking, Vol. 5, No. 6, pp. 835-846, 1997.

[5] Vern Paxson and Sally Floyd. Wide-Area Traffic: The Failure of Poisson Modeling, IEEE/ACM Transactions on Networking, Vol. 3, No. 3, pp. 226-244, 1995.

[6] W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson. Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level, IEEE/ACM Transactions on Networking, Vol. 5, No. 1, pp. 71-86, 2002.

[7] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (extended version), IEEE/ACM Transactions on Networking, Vol. 2, No. 1, pp. 1-15, 1994.

[8] R.Jaiswal, A.Bakre, A.Gutte. Performance Analysis of IPv4 and IPv6 Internet traffic, IJCT, pp.1208-15, 2015.

[9] D.Erman, D.Ilie, A.Popescu. Measuring And Modeling the BitTorrent contention distribution system from Computer Communications, ScienceDirect, pp. 22-29,2010.

[10] A.Legout, G.Urvoy-Keller, and P.Michiardi. Rarest first and choke algorithms are enough, in IMC '06: Proceedings of the 6th ACM SIGCOMM on Internet *measurement*, New York, NY, USA, 2006, pp.203–216, ACM Press.

[11] Daniel Stutzbach and Reza Rejaie. Understanding Churn in Peer-to-Peer Networks, in IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet Measurement, New York, NY, USA, 2006, pp.189–202, ACM.

[12] J.Chandra, N.Ganguly.Optimizing Topology in BitTorrent Based Networks, Workshop on Network Science For Communication Networks,IEEE, pp.888-893.2011.

[13] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, X. Zhang,"Measurements, analysis, and modeling of BitTorrent-like" in Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, 2005: 155-169.

[14] Adar, E., AND B.Huberman. Free riding on gnutella. Tech. rep. , Xerox PARC , August 2000.

[15] Mark E. Crovella and Murad S. Taqqu. Estimating the heavy tail index from scaling properties. Methodology and Computing in Applied Probability, Vol 1(No. 1), 1999.

[16] M. Izal, G. Urvoy-Keller, E.W. Biersack, P.A. Felber, A. Al Hamra, and L. Garc´es-Erice. "Dissecting BitTorrent: Five months in a torrent's lifetime". In *PAM2004*, 2004.

[17] Qiu D. and Srikant R.J. Modeling and performance analysis of bittorrentlike peer-to-peer networks. Technical report, University of Illinois at Urbana-Champaign, USA, 2004.

[18] Thomas Karagiannis, Andre Broido, Michalis Faloustos, and Claffy Kc. Transport layer identification of P2P traffic.IMC'04, 2004.

[19] C.Walck. Handbook for Statistical Distributions For the Experimentalists,Internal Report, Stockholm, 2007.

[20] D.Manini, R.Gaeta. Performance modeling Of P2P file, Workshop on Techniques for processing Complex Sysytems,IEEE,2005.

[21] R.Aristizabal. Estimating The Parameters of the three-parameter Log-Normal Distribution, FIU Digital Commons,2012

[22] D.Erman.2005. BitTorrent Traffic Measurements and Models, Dept .of Telecommunication, Master Thesis,Blekinge Institute of Technology.

[23] K.Park, G. Kim and M.Crovella. On the Relationship between File Sizes, Transport Protocols, and Self-Similar Network Traffic, Proceedings of the 4th International Conference on Network Protocols, pp. 171-180, 1996.