Survey on Recent Applications of Extreme Value Theory in Networking

Jana Nina Friedrich, Max Helm*
*Chair of Network Architectures and Services
School of Computation, Information and Technology, Technical University of Munich, Germany
Email: jana.friedrich@tum.de, helm@net.in.tum.de

***Abstract*—The statistical model Extreme Value Theory (EVT) predicts extreme events, e.g., extreme latencies in networking. This paper summarizes recent applications of EVT in networking from 2022 to 2023 to provide an overview of the current state of the field. The selected nine papers cover the application areas of Flow-Level Tail Latency, Ultra-Reliable Low Latency Communication, Dynamic Service Chaining, Mobile Edge Computing and Root Cause Location. EVT is a powerful method to improve various services, especially for those who need ultra-reliability. However, EVT has limitations. The biggest ones are that the quality of EVT depends on the data volume used, the confidence level employed in the distribution fitting, and the approach used for return level calculation.**

***Index Terms*—extreme value theory, recent applications in networking, literature review, survey**

1. Introduction

Extreme Value Theory (EVT) is a statistical method typically used to predict extreme events and model the data's tail behavior. Its application areas are wide-ranging, from natural catastrophes to engineering. Recent applications, especially in networking, are of interest due to their considerable potential for utilization. Therefore, this paper summarizes the most relevant papers from 2022 to 2023. In order to estimate which works are relevant, the citation and viewing numbers, as well as the standing of the publisher, were taken into account. Papers that do not actively incorporate EVT in their methodology, such as those employing it merely for validating their proposed systems (e.g., Chaccour et al. [1]) or relying solely on an EVT-based model (e.g., Pan et al. [2]), are excluded from consideration. This paper first introduces EVT to establish a foundational understanding and then presents the summaries of the selected papers. The papers cover the application areas of Flow-Level Tail Latency (Chapter 3.1.), Ultra-Reliable Low Latency Communication (Chapter 3.2.), Dynamic Service Chaining (Chapter 3.3.), Mobile Edge Computing (Chapter 3.4.), and Root Cause Location (Chapter 3.5.). Some application areas summarize more than one paper. In the end, the paper concludes with its own take on the field.

2. Understanding Extreme Value Theory

As described by Coles in [3], EVT has grown into an essential statistical model for applied sciences over the last few years. EVT models the tail distribution of empirically collected data and can even be used to predict future extreme events. The characteristic feature of extreme value analysis aims to quantify the stochastic behavior of a process at extremely large or small levels. The extreme value analysis typically requires an estimate of the probability of extreme events that surpass the already observed events. The following subsections present two commonly used distribution approaches of EVT.

2.1. Generalized Extreme Value Distribution

The Generalized Extreme Value (GEV) distribution combines the distribution families of Gumbel, Fréchet and Weibull. Using the combination is more effective than computing which distribution fits the dataset the best. The combined distribution is shown in Equation (1). [3]

$$G(z) = \exp\left\{-\left[1+\epsilon\left(\frac{z-\mu}{\sigma}\right)^{\frac{-1}{\epsilon}}\right]\right\} \qquad (1)$$

Equation (1) has three parameters. μ describes the location, σ the scale and ϵ the tail. By analyzing ϵ through inference, the data autonomously identifies the most suitable tail behavior, eliminating the need for subjective a priori judgments regarding the adoption of a specific extreme value family. Additionally, the uncertainty in the inferred value of ϵ quantifies the lack of certainty about which of the original three distribution families is most appropriate for a given dataset. GEV is used to model the distribution of block maxima. It separates the data into blocks of the same length and fits the GEV to the resulting set of block maxima. The choice of block size becomes critical when applying this model to any dataset. This decision involves balancing bias and variance. Smaller blocks may lead to poor model approximation, while larger blocks increase estimation variance. There are different methods to estimate the parameters of GEV. The most common is likelihood-based. However, by employing the likelihood-based method, a challenge arises around the regularity condition. This condition is essential for ensuring the validity of typical asymptotic properties linked to the maximum likelihood estimator. This challenge emerges from the GEV model because the endpoints of the distribution are functions of the parameter values. In the next subsection, another EVT approach solves this problem. [3]

2.2. Generalized Pareto Distribution

Focusing solely on modeling block maxima is an inefficient approach to extreme value analysis when additional

 doi: 10.2313/NET-2024-04-1_11

data on extremes is accessible. [3]

As described by Haan et al. [4], the Generalized Pareto Distribution (GPD) uses the Peaks over Threshold (PoT) approach instead. PoT categorizes all data points surpassing a chosen threshold as part of the tail. Equation (2) and (3) describe the GPD approach.

$$H(z) = 1 - \left(1 + \frac{\epsilon z}{x}\right)^{\frac{-1}{\epsilon}} \tag{2}$$

$$x = \sigma + \epsilon\,(y - \mu) \tag{3}$$

GPD has the same three parameters as GEV. y is the selected threshold. Altering the block size, even if it remains large, would impact the GEV parameters but not GPD. ϵ remains constant to block size changes, and the computation of x in Eq. (3) also remains unaffected. Variations in μ and σ are self-compensating. The GPD distribution, once fitted, has various applications. One possibility is to compute the return level corresponding to a given return period. This calculated value represents the extreme event. On average, an extreme event occurs once during that period. [3]

3. Recent Applications in Networking

The following subsections summarize nine different papers. Each summary consists of, when deemed necessary, a brief introduction to the topic, followed by an exposition of the proposed contributions of the discussed paper. Then, this paper explains the methodology used to make these contributions and presents the research results. All nine papers apply the EVT as shown in Figure 1. The papers first collect and filter data and then assess whether the collected data is suitable for the application of the EVT. Subsequent subsections discuss the criteria for applying the EVT. The papers gather additional data if the criteria are not met. On the other hand, if they are satisfied, the parameters of the EVT (such as the threshold value for the GPD approach) are calculated. Finally, the papers validate and apply their EVT model.

3.1. Flow-Level Tail Latency

This subsection summarizes the paper of Helm et al. [5].

Requirements at the end-to-end latency can be used in service-level agreements for communication networks and can, therefore, influence network planning and flow admission. These latencies can be measured and used as input for models, like EVT, to predict extreme latency occurings. The paper uses the PoT approach for 100 networks with random topologies, flow specifics and configurations to show that EVT can be applied to large datasets. The authors use 14 billion latency and jitter values from the measurements of Wiedner et al. [6]. Then, the EVT model is derived from the first 5 % of the data and validated on the remaining 95 %. Flow-level models outperform network-level models for high percentiles, suggesting that EVT models are more suitable at the flow-level when focusing on high percentiles of the tail. In addition, these models have a lower relative error of percentile values compared to network-level models. This result indicates their superior suitability despite having

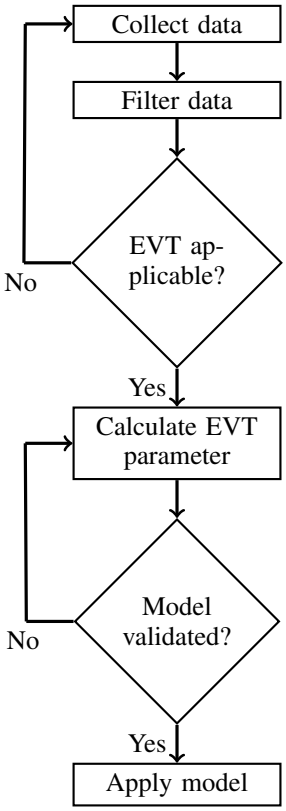

Figure 1: Overview flow of how to use EVT

less data. The accuracy of predictions ranges from 75 % to 85 % for twenty- and twofold time horizons. Two times the time refers to a duration corresponding to two times of the given horizon. The calculation is similar for twenty times. The paper forecasts tail latency quantiles at the flow-level with median absolute percentage errors between 0.7 % to 16.8 %. However, there are limitations to EVT. It depends on the volume of data, the confidence level of the distribution fitting, and the return level calculation. In addition, EVT is only applicable if the data is identically distributed and stationary. The authors use the Augmented Dickey-Fuller (ADF) test to ensure stationary. While in their setup, most flow latencies are stationary, this is not a general assumption.

3.2. Ultra-Reliable Low Latency Communications

This subsection summarizes the four papers from Mehrnia et al. [7]–[10]. These papers build on each other and use previous proofed results.

Ultra-Reliable Low Latency Communications (URLLC) is vital for 5th generation communication networks. An accurate channel modeling is needed since URLLC has a strict packet error rate and latency requirements. The paper [7] introduces a wireless channel modeling methodology based on EVT. The methodology involves deriving the parameters of the tail distribution by fitting the GPD to independent and identically distributed (i.i.d.) samples. To obtain these samples, the authors use declustering methods, Auto-Regressive Integrated Moving Average, and Generalized

Auto-Regressive Conditional Heteroskedasticity. After applying EVT, the mean residual life and parameter stability methods determine the optimum threshold. Next, the algorithm Minimum Sample Size Determination specifies the stopping condition to calculate the minimum required number of samples. Lastly, probability plots such as Probability/Probability (PP) and Quantile/Quantile (QQ) validate the channel tail model. The proposed framework requires significantly fewer samples than the conventional extrapolation-based approach. In addition, it fits the empirical data in the lower tail better.

The authors in [8] introduce a framework based on EVT to compute the optimal transmission rate in Ultra-Reliable Communication (URC). The authors consider a URC system that encounters fading, leading to diminished received power values. The system consists of a transmitter and receiver, sending packets over an unknown stationary channel. First, GPD represents the channel. The received power samples convert through declustering to i.i.d samples. These samples fit the GPD to the lower tail, and PP and QQ plots validate the Pareto model. Next, the optimal transmission rate is estimated. The function that selects the rate uses Pareto parameters. Finally, an evaluation of the error probability verifies the selected rate. The proposed framework outperforms traditional methods regarding reliability. Traditional methods use average statistic channel models. Moreover, because of the usage of the GPD threshold, the number of samples needed to achieve a certain reliability could be minimized.

In [9], a novel EVT-based framework is proposed to estimate the optimal transmission rate as well as the confidence interval on a small number of samples. The system model consists of one transmitter and one receiver communicating over a channel. The ADF test checks if the channel is stationary. If not, all factors that cause time variation of the GPD parameters are determined and collected as a sequence. This sequence is split into as many groups as needed, making each group stationary. The fixed transmission power is known in advance. First, the transmitter sends a packet to the receiver over an unknown channel. Then, the tail distribution is estimated by applying a modeling methodology based on EVT on the channel. Therefore, the statistics fit GPD to obtain power values surpassing the provided threshold. Moreover, the confidence intervals of wrong conclusions are derived for various numbers of samples. The intervals correspond to different probabilities. Lastly, the paper assesses the transmission rate by using EVT. Thus, the intervals of the Pareto parameters from different sample sizes are incorporated to achieve the desired error probability in URC. The paper validates its proposed framework with data collected in different sizes in a car engine. The targeted error probability is even met with limited data known.

As described in [10], the statistical method Multivariate Extreme Value Theory (MEVT) models the relation of rare events based on multidimensional limiting relations. MEVT is a further development of EVT and has additional functions like modeling dependence structures and joint distribution of several extreme events. The authors of [10] base their proposed channel modeling methodology on MEVT. The channel is for systems using Multiple Input Multiple Output (MIMO)-URC for efficiently deriving the lower tail statistic in multiple dimensions. The received signal powers are the data for these statistics. To validate the proposed methodology, the paper focuses on the bivariante or two-dimensional case. Before MEVT can be applied, the collected data converts into a sequence of i.i.d samples. Therefore, the above-introduced modeling of paper [7] is applied. Next, MEVT fits the GPD to the tail distribution to find optimal thresholds. Afterward, the Fréchet transformation is applied to each data sequence. Then, between the Fréchet sequences, the dependency factor is estimated. Next, two approaches are employed to fit Bi-Variate GPD (BGPD) to the joint distribution. The approaches used are the logistical distribution and the Poisson point process. Lastly, a mean constraint assessment validates the fitted BGPD model. The methodology is tested using one transmitter and two receivers in a car engine and compared to conventional models based on extrapolation. The proposed method performs significantly better in accurately modeling multiple dimensions events in URC.

3.3. Dynamic Service Chaining

This subsection summarizes the paper of Qin et al. [11].

Physical Machines (PMs) host Virtual Machines (VMs). VMs run software-based Virtual Network Functions (VNFs), which are enabled by Network Function Virtualizations (NFVs). The most essential requirement for service function chaining is guaranteeing ultra-reliable services. The existing research concentrates on average inter-failure time and repair downtime to define the reliability of VNFs. Due to uncertain PM failures, this does not fully capture the stochastic nature of VNF failure. The paper proposes a Dynamic Service Chaining (DSC) framework to examine the high-order statistics and probability of VNF failure time threshold deviation. The GPD approach of EVT characterizes the threshold deviation statistics with a low occurrence probability. The Poisson-Bernstein de la Harpe (PBdH) theorem describes extreme cases of PM failure time. A two-timescale VNF framework for mapping/remapping handles uncertain PM failure. The primary remapping framework works at a large timescale using matching theory. The optimal backup VNF framework operates at a smaller timescale. The algorithm used to find this backup effectively reduces computational complexity and balances switching costs and reliability. In addition, the backup needs to be selected beforehand. Simulation of the proposed DSC validates the PBdH. The randomly generated network topology is based on 20 nodes and 40 links. Other parameters are normalized. The numerical results show that using EVT to characterize extreme events improves service reliability compared to average-based schemes.

3.4. Mobile Edge Computing

This subsection summarizes first the paper of Liu et al. [12] and then of Ji et al. [13].

Traditional cloud computing has its resources pooled centrally. Mobile Edge Computing (MEC) has an advantage compared to traditional approaches because it

provides computing services close to the server. The authors address the challenges of offloading mission-critical tasks in MEC networks with Non-Orthogonal Multiple Access (NOMA). The network of the paper consists of a server and two sensor nodes, which supply the server with data. The server computes latency-sensitive tasks and works after the first-come, first-served principle. The overall error probability is characterized by the derivation of the Finite Blocklength (FBL) communication reliability and latency violation error probability through the GPD approach of EVT. The framework minimizes errors by jointly allocating the communication phase, the computation phase, and the user transmits power within stringent delay and energy constraints. The modified Block Coordinate Descent method addresses the non-convex problem by optimizing the time duration or proving the problem by characterizing the joint convexity of FBL error probability. Numerical simulations confirm the near-optimal performance of the proposed approach. Moreover, the paper's proposed framework outperforms the NOMA scheme with infinite blocklength solutions and the time-division multiple access scheme.

The authors in [13] address the issue of energy-efficient computation offloading in MEC systems on mobile applications with sequential or parallel module dependencies. First, the authors model mobile applications as Directed Acyclic Graphs. By considering the parent and children set of each computation module, the execution dependency gets handled. Then, the GEV approach of EVT is applied to explicitly address uncertainties and limit the occurrence probability of extreme events. Afterward, a newly developed ϵ-bounded algorithm, based on the column generation technique and with theoretical optimality guarantees, solves the offloading problem energy-efficiently. ϵ is the tail of the GEV model, and the optimal offloading policy is when ϵ is 0. Tools like Smart Diagnoses, tPacketCapture, WiFi SNR and PETrA were used to measure and record statistics. The result is a computation scheme outperforming other state-of-the-art schemes, such as Hermes and JSCO, in experiments conducted on an Android platform. The proposed scheme consistently has the lowest energy consumption as long as ϵ is smaller than 0.05. When this happens, the local device can save up to $50\,\%$ of energy. The cause for this is that JSCO neglects to account for uncertainties inherent in dynamic radio channels with queueing delays. In contrast, Hermes introduces additional energy consumption attributed to communication overhead resulting from the continuous probing of the channel.

3.5. Root Cause Location

This subsection summarizes the paper of Yang et al. [14].

The increasing complexity of online services can lead to significant losses when abnormalities occur. The root cause location is vital to guarantee the stable operation of online services. Therefore, the paper proposes a location method based on Prophet and Kernel Density Estimation (ProphetKdeRCL). ProphetKdeRCL consists of two stages. The first stage is the abnormal detection of performance indicators. This stage introduces the Prophet Mutation Point Updating (PMPU) algorithm. The Prophet model fits trend items better, and the timing anomaly detection gets more accurate through the usage of an improved version of EVT. PMPU solves the problems of existing methods since it can detect irregularities in the lowest range. The second stage locates the root cause of abnormal indicators and uses two algorithms. One is an anomaly degree measurement algorithm based on a Kernel Density Estimation. The second one is a time window-based causality analysis algorithm. This algorithm analyzes latency dependency via an intermediate structure and a time window. The effectiveness of the proposed algorithm is validated through testing and evaluations of the public time series, the microservice application system fault detection, and root cause location datasets.

4. Conclusion

Extreme Value Theory is a robust statistical method for predicting occurring extreme events and tail behavior modeling. The focus of this survey paper is on the most recent applications in networking from 2022 to 2023. This paper excludes papers that only build on EVT-based models or EVT for validating the paper's proposed methodology. First, this paper establishes an understanding of EVT. Therefore, it explains the distributions of Generalized Extreme Value and the Generalized Pareto Distribution in detail. Second, the nine selected papers are summarized. Each summary consists of the contribution of the paper, used methodologies, and key findings. Table 1 gives an overview of all discussed papers. Each line represents one of the nine papers. The order is the same as the papers are summarized in this paper. The Approach column shows that EVT's GPD approach is preferred over the GEV approach. All papers validate their EVT-based approaches and evaluate that their approach is superior to traditional ones in the discussed scenarios. They prove that accurate extreme event modeling and improving reliability predictions are possible. Therefore, they test their approach through virtual simulations or in the real world. None of the papers published their data to replicate their tests, and none have a Reproducibility Badge from the Association for Computing Machinery. Nevertheless, EVT has limitations, but only [5] highlights them. The quality of the EVT depends on factors such as the data volume, the confidence level employed in the distribution fitting, and the approach to return level calculation. Another weakness of EVT is that it can only be applied if the data is identically distributed and stationary. The data and used communication channels have to be either tested for stationary or assumed to be stationary, which leads to more calculation effort and complex systems. In conclusion, it can be said that EVT is a powerful tool to model and predict extreme values if the right approach is selected, the data is optimally fitted, and enough meaningful data volume is available. If not, EVT increases the system complexity and delivers wrong predictions. EVT is especially useful in the field of telecommunication since the demand for a method that can handle computation-intensive and latency-critical tasks is met.

TABLE 1: Overview table of all summarized papers

Paper	Approach	Validated	Tested
Helm et al. [5]	GPD	Yes	Virtual
Mehrnia et al. [7]	GPD	Yes	Real-World
Mehrnia et al. [8]	GPD	Yes	Real-World
Mehrnia et al. [9]	GPD	Yes	Real-World
Mehrnia et al. [10]	BGPD	Yes	Real-World
Qin et al. [11]	GPD	Yes	Virtual
Liu et al. [12]	GPD	Yes	Virtual
Ji et al. [13]	GEV	Yes	Real-World
Yang et al. [14]	improved EVT	Yes	Virtual

References

[1] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can terahertz provide high-rate reliable low-latency communications for wireless vr?" *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9712–9729, 2022.

[2] C. Pan, Z. Wang, H. Liao, Z. Zhou, X. Wang, M. Tariq, and S. Al-Otaibi, "Asynchronous federated deep reinforcement learning-based urllc-aware computation offloading in space-assisted vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7377–7389, 2023.

[3] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001.

[4] F. Haan, *Extreme Value Theory*. Springer New York, NY, 2010.

[5] M. Helm, F. Wiedner, and G. Carle, "Flow-level tail latency estimation and verification based on extreme value theory," in *2022 18th International Conference on Network and Service Management (CNSM)*, 2022, pp. 359–363.

[6] F. Wiedner, M. Helm, S. Gallenmüller, and G. Carle, "Hvnet: Hardware-assisted virtual networking on a single physical host," 2022.

[7] N. Mehrnia and S. Coleri, "Wireless channel modeling based on extreme value theory for ultra-reliable communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 1064–1076, 2022.

[8] ——, "Extreme value theory based rate selection for ultra-reliable communications," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6727–6731, 2022.

[9] ——, "Incorporation of confidence interval into rate selection based on the extreme value theory for ultra-reliable communications," in *2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2022, pp. 118–123.

[10] ——, "Multivariate extreme value theory based channel modeling for ultra-reliable communications," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[11] S. Qin, M. Liu, and G. Feng, "Dynamic service chaining for ultra-reliable services in softwarized networks," *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 3585–3595, 2023.

[12] Z. Liu, Y. Zhu, Y. Hu, P. Sun, and A. Schmeink, "Reliability-oriented design framework in noma-assisted mobile edge computing," *IEEE Access*, vol. 10, pp. 103 598–103 609, 2022.

[13] T. Ji, C. Luo, L. Yu, Q. Wang, S. Chen, A. Thapa, and P. Li, "Energy-efficient computation offloading in mobile edge computing systems with uncertainties," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 5717–5729, 2022.

[14] Y. Yang, Y. Sun, Y. Long, J. Mei, and P. Yu, "Root cause location based on prophet and kernel density estimation," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 904–917, 2023.