Modelling Traffic Distributions Using Probabilistic Machine Learning and Extreme Value Theory

Akanksha Ahuja*, Sam Nallaperuma-Herzberg*, Albert Rafel†, Andrew Lord†, Seb J. Savory*
*University of Cambridge, CB3 0FA, UK, †British Telecommunications, Ipswich, IP5 3RE, UK
aa2419@cam.ac.uk

***Abstract*—Traffic on national optical backbones is strongly influenced by rare, high-demand flows that shape capacity dimensioning and resilience assessment. However, due to the sensitive nature of commercial traffic data, comprehensive origin–destination measurements are rarely available. As a result, traffic models used in network planning commonly rely on heuristic demand formulations or simplified distributional assumptions calibrated to average behaviour. In practice, these approaches under-represent extreme demand events, leading to biased estimates of peak load and stress conditions that drive design and upgrade decisions. In this work, we propose a hybrid distributional traffic modelling framework that addresses this limitation by combining Gaussian Mixture Models (GMM) with Extreme Value Theory (EVT). Traffic demand is decoupled into bulk and tail regimes, with GMM capturing typical demand patterns and EVT modelling rare, high-impact extremes relevant for planning. We show that the proposed approach reproduces observed traffic distributions more accurately than commonly used heuristic, parametric, and supervised baselines, achieving an order-of-magnitude reduction in error. In particular, EVT-based tail modelling reduces error by approximately 40% relative to the lognormal distribution, which is widely adopted in practice. We validate the framework using real traffic from the British Telecom (UK) national backbone. The resulting dual-regime model is interpretable, and enables realistic, privacy-preserving generation of traffic for stress testing in data-poor planning settings.**

Keywords**-** Traffic Distributions, Gaussian Mixture Models, Extreme Value Theory, Privacy-preserving, Generative

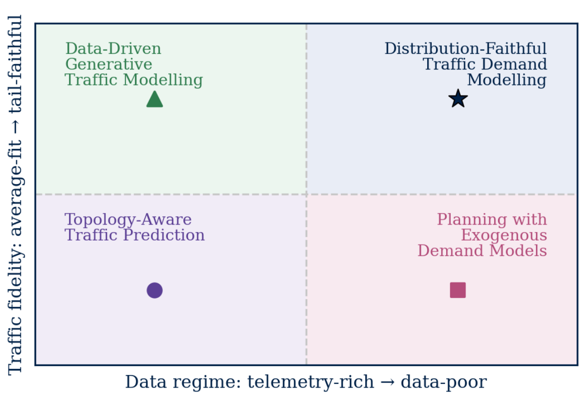

Fig. 1. The Traffic Modelling Landscape. Current approaches force a trade-off between Data Availability and Distributional Fidelity. Predictive and Generative models (Left) require dense telemetry, while Planning heuristics (Bottom-Right) rely on simplified averages. This work (Star, Top-Right) bridges the gap, offering high tail fidelity even in data-scarce regimes.

I. Introduction: Traffic Distribution Modelling

Network failures and blocking events are driven by rare, high-demand extremes. When such tail effects are not adequately represented, the resulting designs may appear optimal under nominal loads but remain vulnerable to congestion during peak demand. Therefore, inaccuracies in traffic modelling propagate into planning decisions, and introduce latent risk into the infrastructure. Existing predictive models [1]–[4] and deep generative approaches [5] are not explainable and rely on rich datasets (Figure 1). These approaches cannot be used in data-scarce planning scenarios. Conversely, traditional network planning frameworks treat demand as an exogenous input derived from heuristics [6], [7], which do not capture the real-world traffic patterns. Furthermore, distributional approaches enforce single-family assumptions that fit the bulk well but underestimate the extreme tails [8], [9]. To address this gap between data efficiency and tail fidelity, we propose a hybrid GMM-EVT framework to explicitly decouple and model these distinct traffic regimes, in an interpretable manner.

II. Method: Generative Model with Machine Learning and Extreme Value Statistics

We model traffic demand using a two-regime parametric formulation that separates bulk and extreme behaviour (Figure 2). This results in a hybrid generative framework that combines Gaussian Mixture Models (GMMs) with Extreme Value Theory (EVT) [9], [10]. Let $X \in [0, 1]$ denote the normalized, aggregated bidirectional flow between node pairs in a network $G(V, E)$.

The empirical distribution $p_{\text{data}}(x)$ is partitioned into two regimes by introducing a threshold u, defined as the empirical 95^{th} percentile of the observed traffic consistent with prior studies [8]. This threshold ensures the asymptotic validity of extreme models and enables fair cross-topology comparisons.

We model the bulk ($x \leq u$) and tail ($x > u$) separately using parametric models. This results in a piecewise generative formulation,

$$f(x \mid \Theta) = \begin{cases} f_{\text{bulk}}(x \mid \Theta_{\text{bulk}}), & x \leq u, \\ f_{\text{tail}}(x \mid \Theta_{\text{tail}}), & x > u, \end{cases} \tag{1}$$

where Θ_{bulk} and Θ_{tail} denote the parameters of the bulk and tail models, respectively. The bulk traffic set $\mathcal{X}_{\text{bulk}} = \{x \in \mathcal{X} \mid x \leq u\}$ captures the bulk traffic regime. We model the bulk density using a Gaussian Mixture Model (GMM), expressed as a weighted sum of K Gaussian components:

$$f_{\text{bulk}}(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x \mid \mu_k, \sigma_k^2), \tag{2}$$

978-3-903176-78-2

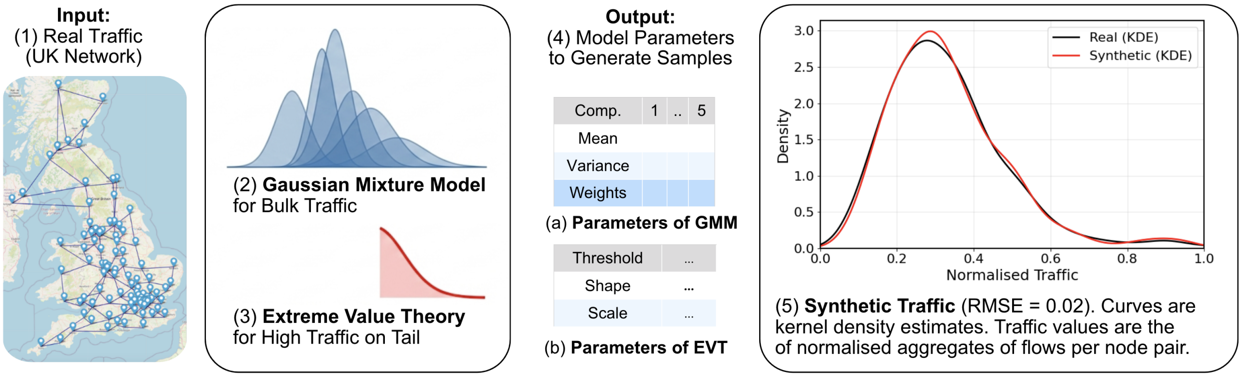

Fig. 2. Overview of the pipeline for traffic distribution modelling and its ability to capture both bulk and extreme behaviour in networks. Real traffic (325 node pairs) from the UK core network is decomposed into a bulk regime modelled using a Gaussian Mixture Model and a tail regime modelled using Extreme Value Theory. The learned parameters jointly generate synthetic traffic that matches the empirical distribution, as shown by the aligned kernel density estimates.

TABLE I
COMPARISON OF SYNTHETIC TRAFFIC GENERATION AND PREDICTION MODELS. LOWER VALUES INDICATE BETTER PERFORMANCE.

Model	KL Divergence ↓	MSE ↓	RMSE ↓
GMM + EVT (Proposed)	**0.0012**	**0.0003**	**0.0159**
GMM only	0.0017	0.0004	0.0212
Lognormal	0.0019	0.0013	0.0365
Gamma	0.0025	0.0012	0.0342
Beta	0.0079	0.0035	0.0590
Normal	0.0150	0.0025	0.0497
Uniform	0.0574	0.0735	0.2712
EVT only	0.1163	0.0101	0.1007
Mean	0.1003	0.0235	0.1532
Median	0.1003	0.0236	0.1536
Linear Regression	0.0144	0.0035	0.0593
Random Forest	0.0234	0.0049	0.0698
Gravity Model	16.6775	0.1292	0.3594
DC-IXP Model	10.7277	0.1145	0.3383

where π_k are mixture weights satisfying $\sum_{k=1}^{K} \pi_k = 1$, and (μ_k, σ_k^2) denote the mean and variance of the k-th component. The parameter set is $\Theta_{\text{bulk}} = \{\pi_k, \mu_k, \sigma_k^2\}_{k=1}^{K}$. The parameters Θ_{bulk} are estimated via the Expectation–Maximisation (EM) algorithm, which iteratively maximises the log-likelihood of the bulk data $\mathcal{X}_{\text{bulk}}$. For the upper tail region ($x > u$), standard parametric distributions are not able to capture the decay associated with extreme congestion events. We therefore apply Extreme Value Theory (EVT) using the Peaks-Over-Threshold (POT) framework. The exceedances $Y = X - u$, are modelled using the Generalised Pareto Distribution (GPD), whose density is given by

$$f_{\text{tail}}(y \mid \xi, \sigma) = \frac{1}{\sigma}\left(1 + \xi \frac{y}{\sigma}\right)^{-\left(1+\frac{1}{\xi}\right)}, \tag{3}$$

where ξ is the shape parameter and $\sigma > 0$ is the scale parameter. The tail parameters $\Theta_{\text{tail}} = \{\xi, \sigma\}$ are estimated via Maximum Likelihood Estimation (MLE) using the exceedance set $\mathcal{X}_{\text{tail}} = \{x - u \mid x > u\}$.

Synthetic traffic is generated by stratified sampling that preserves the empirical proportion of extreme events. Samples are drawn from the fitted Gaussian mixture for the bulk regime and from the Generalised Pareto distribution for exceedances above u, and then combined to form a synthetic traffic distribution.

TABLE II
GAUSSIAN MIXTURE MODEL PARAMETERS

Component	Mean	Variance	Weight
1	0.2065	0.000689	0.2407
2	0.4958	0.002656	0.1757
3	0.3782	0.000646	0.2061
4	0.1250	0.000530	0.0991
5	0.2922	0.000686	0.2785

TABLE III
GENERALISED PARETO DISTRIBUTION (GPD) PARAMETERS

Shape (ξ)	Location	Scale	Threshold u
-0.8663	0.0	0.3581	0.5917

III. EVALUATION AND DISCUSSION

The objective is faithful reproduction of traffic distributions used in planning studies. We evaluate the proposed GMM–EVT framework against classical distributional models, heuristic planners, and topology-aware regression baselines using both distributional metrics (Kullback–Leibler (KL) divergence) and error-based metrics (mean squared error (MSE) and root mean squared error (RMSE)) (Table I). RMSE and MSE are computed on sorted samples to quantify distributional mismatch rather than pointwise prediction error. This evaluation assesses the realism of synthetic traffic.

Performance: Figure 2 summarises the GMM–EVT pipeline and its distributional fidelity, illustrated by overlapping kernel density estimate (KDE) curves of real and synthetic traffic. GMM–EVT achieves the lowest error (RMSE $= 0.016$), outperforming the lognormal distribution (0.037), the state-of-the-art model for Internet traffic [8] (Table I). Figure 3 uses log-scale complementary cumulative distribution functions (CCDFs), representing $P(X \geq x)$, to assess how well the model reproduces tail behaviour and the transition from bulk to extreme demand. Agreement is observed in the tail region, where the real and synthetic curves assign similar probabilities to high traffic values. For example, at $x \approx 0.75$, both curves give $P(X \geq x) \approx 10^{-2}$. This indicates that the model preserves the rarity of high-demand events, which is not captured by standard traffic models. Minor deviations at traffic

TABLE IV
RMSE COMPARISON ACROSS BULK AND TAIL REGIMES

Network	Regime	Lognormal ↓	GMM–EVT ↓	Gain ↑
UK	Bulk	0.0349	**0.0143**	59.0%
	Tail	0.0601	**0.0355**	40.9%
	Overall	0.0365	**0.0159**	56.4%
TEFNET	Bulk	0.0137	**0.0067**	51.1%
	Tail	0.1624	**0.1135**	30.1%
	Overall	0.0396	**0.0269**	32.1%

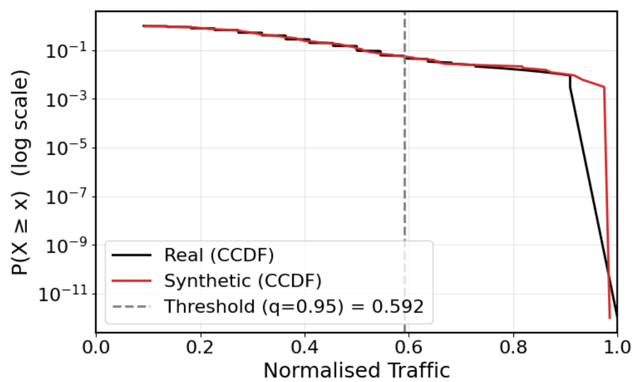

Fig. 3. Log-scale CCDF comparison of real and synthetic traffic to assess their agreement in the tail. The CCDF represents the probability that traffic exceeds a given value. The dashed line marks the threshold ($q = 0.95$), which separates the bulk regime (left) from the tail regime (right).

values above 0.9 are attributable to the sparsity of the far-tail, where the exceedance set contains only a few high-magnitude samples to support parametric estimation. Domain-specific heuristics perform substantially worse as gravity models (0.35) [7], data-centre and peering heuristics (0.33) [6], and uniform distributions (0.27) incur errors an order of magnitude higher.

Comparison with Topological Learning: We further investigate whether contextual features improve distributional fidelity and evaluate supervised baselines that are widely adopted by operators (linear regression, random forests, [3]) using multiple features such as geolocation, population, data-centre and Internet Exchange Point features, and node degrees. While these methods outperform heuristics (RMSE $\approx$ 0.06–0.07), they still fail to match GMM–EVT's performance.

Ablations: Sensitivity analysis over the number of Gaussian components (K = 1-19) shows that RMSE improves up to K = 5, beyond which no further gains are observed for the tail RMSE. Sensitivity analysis over threshold selection (75–95%) indicates that tail RMSE is minimised around an 85% threshold, with $K = 5$ Gaussian components. Finally, Table I shows that the unified GMM–EVT formulation consistently outperforms both GMM-only and EVT-only variants.

Interpretability: In contrast to implicit generative models such as generative adversarial networks (GANs), variational autoencoders (VAEs) and diffusion models, which define distributions through sampling [5], the proposed model is fully parametric. The learned parameters (Tables II and III) are directly interpretable and sufficient to generate realistic traffic without revealing source–destination flows.

Generalisation: We test whether the same 5-component GMM–EVT model is transferable to another network (TEFNET) [11]. GMM bulk parameters ($\mu \pm \sigma$) shift from 0.309 ± 0.119 (UK) to 0.055 ± 0.075 (TEFNET). Concurrently, the EVT threshold decreases from $u = 0.592$ to 0.347, with the shape parameter ξ becoming more negative (-0.866 to -1.483). These shifts confirm the framework's ability to adapt to distinct traffic distributions while preserving a fixed, interpretable model structure. Finally, our model reduces tail estimation error by approximately 30-40% relative to the lognormal model for both networks (Table IV).

IV. CONCLUSION: REALISTIC TRAFFIC MODELLING

We introduce a statistical–learning framework for modelling optical network traffic under limited data. Gaussian Mixture Models combined with Extreme Value Theory capture the multi-modal bulk (low and moderate traffic) and the tail (high traffic) across the UK optical network. (i) GMM–EVT reduces global error by an order of magnitude compared to heuristic baselines (10-20 times). (ii) It further improves extreme traffic estimation by approximately 40% relative to the industry-standard. (iii) Supervised learning approaches even with topological and spatial features severely underperform (3-4 times worse) in this data-scarce regime. (iv) Our results suggest that the proposed framework is generalisable to other large-scale backbone networks. (v) It supports privacy-preserving sharing with only 18 distributional values per network (15 GMM parameters, 2 EVT parameters, and the threshold u). This work models traffic as static aggregated demand. Future work will extend the framework to include temporal traffic, address the spatial assignment of generated demands to specific node pairs, and integrate tail-risk profiles into resilience analysis.

Acknowledgements: This work was supported by the EPSRC Programme Grant TRANSNET [Grant Number EP/R035342/1] and an EPSRC / BT ICASE Studentship. The authors would like to acknowledge Paul Wright for kindly providing the dataset used in this work.

REFERENCES

[1] G. Choudhury, D. Lynch, G. Thakur, and S. Tse, "*Two Use Cases of ML for SDN-Enabled IP/Optical Networks* ," *JOCN*, 2018.

[2] M. Aibin, N. Chung, T. Gordon, L. Lyford, and C. Vinchoff, "*On Short- and Long-Term Traffic Prediction in Opt. Net. Using ML*," *ONDM*, 2021.

[3] D. Szostak, A. Włodarczyk, and K. Walkowiak, "*ML Classification and Regression Approaches for Opt. Net. Traffic Pred.*," *Electronics*, 2021.

[4] B. Bao, *et al.* "*Node-Oriented Traffic Prediction and Scheduling Based on GCN in Metro Optical Networks* ," *OFC*, 2021.

[5] A. L. García Navarro, *et al.* "*CTGANs for Generating Synthetic Traffic Matrices in WANs*," *ONDM*, 2025.

[6] S. K. Patri, A. Autenrieth, J. P. Elbers, and C. M. Machuca, "*Planning Optical Networks for Unexpected Traffic Growth*," *ECOC*, 2020.

[7] A. Dwivedi and R. E. Wagner, "*Traffic Model for USA Long-Distance Optical Network*," *OFC*, 2000.

[8] M. Alasmar, R. Clegg, N. Zakhleniuk, and G. Parisis, "*Internet Traffic Volumes Are Not Gaussian—They Are Log-Normal: An 18-Year Longitudinal Study with Implications for Modelling and Prediction* ," *IEEE/ACM Trans. Netw*, 2021.

[9] S. J. Savory, "*Extreme Values in Optical Fiber Comm. Sys.*," *OFC*, 2020.

[10] S. Sahu and J. C. Clement, "*GMM for Enhancing the QoT Estimation in Optical Networks: A Machine Learning Approach* ," *Sci. Rep.*, 2025.

[11] J. M. Rivas-Moscoso, *et al.*, "*TEFNET24: Reference Packet Optical Network Topology for edge to core transport*," *JOCN*, 2024.