Likelihood-Based Inference for Modelling Packet Transit From Thinned Flow Summaries

Prosha Rahman, Boris Beranger, Scott Sisson, and Matthew Roughan, *Fellow, IEEE*

***Abstract*—Network traffic speeds and volumes present practical challenges to analysis. Packet thinning and flow aggregation protocols provide smaller structured data summaries, but conversely impede statistical inference. Methods which model traffic propagation typically do not account for the packet thinning and aggregation in their analysis and are of limited practical use. We introduce a likelihood-based analysis which fully incorporates packet thinning and flow aggregation. Inferences can hence be made for models on the level of individual packets while only observing thinned flow summaries. We establish consistency of the resulting maximum likelihood estimator, derive bounds on the volume of traffic which should be observed to achieve a desired degree of efficiency, and identify an ideal family of models. The robust performance of the estimator is examined through simulated analyses and an application on a publicly accessible trace which captured in excess of 36 m packets over a 1 minute period.**

***Index Terms*—Network analysis, NetFlow, Flow aggregation, Traffic sampling, Symbolic Data Analysis.**

I. Introduction

Network traffic volumes and speeds have grown exponentially since the inception of the internet [1]. Recording and analysing such volumes is impractical and frequently computationally infeasible. Accounting compromises such as *flow aggregation* and *packet thinning* are typically employed to mitigate such volume [2].

Internet traffic is comprised of *packets* which are distributed amongst *flows*. Packets are small quantums of information which, when grouped with other relevant packets, form digital objects. The temporal sequence of a group of packets is called a flow. Flow aggregation is a concession whereby broad flow characteristics — such as the flow size, start time, and duration — are recorded instead of individual packet information. Commonly used protocols include IPFIX and *NetFlow* (here we use NetFlow as a general term to denote flow aggregates). NetFlows reduce the volume of information as each flow is summarised into typically eight numbers [2].

Packet thinning is a parallel strategy whereby only select packets are recorded. Thinning techniques include flow sampling [3], adaptive sampling, and simple random sampling [4]. We shall consider Bernoulli sampling, where packets are independently recorded with some constant known probability q [2]–[4]. In practice, sampling rates can be tuned to match the density and speed of traffic in the network. Flow aggregation can then be analogously applied to thinned traffic.

Such data retention strategies need to be considered explicitly when analysing the summarised data, since bias, and other errors, may otherwise arise. In traffic classification, for example, basic analysis on thinned traffic will over-represent large media applications such as video streaming since these flows are significantly larger. Smaller but more numerous applications, such as e-mails, will conversely be under-represented.

Many of the more sophisticated network analysis techniques exclusively address either flow aggregation [5] or packet thinning, but fail to jointly consider both [3], [6], [7]. Analyses which have mutually addressed packet thinning and flow aggregation have assessed network volume (number of packets in the network) and traffic classification (types of digital objects) [5], [8]. In contrast, we wish to perform parametric inference on patterns of traffic propagation when observing only NetFlows obtained from thinned traffic. More simply, we wish to assess the models which influence the timing of packet arrivals within the network.

In this article, we adapt recent results of [9] and [10] in Symbolic Data Analysis (SDA) to develop a likelihood-based approach for modelling packet-level network traffic. The resulting *NetFlow likelihood* incorporates packet thinning and flow aggregation within a generative framework. As a result, we are able to fit models and make inference on packet-level traffic patterns when observing only flow-level summaries, including those constructed from packet thinned traffic.

We also make three key contributions within the traffic modelling and SDA literature. We first demonstrate that the NetFlow maximum likelihood estimator attains the consistency and asymptotic Normality typical of standard likelihood estimators computed using complete data. These are the first such results for SDA likelihood-based methods. We then provide comparative bounds on the loss of information from the aggregation and thinning procedures. From this, we are able to

Manuscript received 2 December 2021; revised 23 April 2022; accepted 8 June 2022. Date of publication 5 July 2022; date of current version 15 July 2022. This work was supported in part by the Australian Research Council through the Australian Centre of Excellence for Mathematical and Statistical Frontiers under Grant ACEMS, CE140100049, and in part by the Discovery Project Scheme under Grant FT170100079. CAIDA's Internet Traces was supported by the National_Science_Foundation, the U.S. Department of Homeland Security, and CAIDA Members. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Coates. *(Corresponding author: Prosha Rahman.)*

Prosha Rahman, Boris Beranger, and Scott Sisson are with the UNSW Data Science Hub (uDASH), University of New South Wales, Sydney, NSW 2052, Australia (e-mail: p.a.rahman@unsw.edu.au; b.beranger@unsw.edu.au; scott.sisson@unsw.edu.au).

Matthew Roughan is with the School of Mathematical Sciences, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: matthew.roughan@adelaide.edu.au).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSIPN.2022.3188457, provided by the authors.

Digital Object Identifier 10.1109/TSIPN.2022.3188457

identify a lower bound on the minimum number of (packet thinned) flow aggregates required to produce an estimator which approximates the efficiency of the MLE computed on complete data. We also identify a family of models for which inference on the aggregated and complete data are identical. Finally, we introduce an extension of the moments-based estimators of [6] for NetFlow data in order to facilitate comparison with existing approaches.

Despite its many desirable properties, the NetFlow estimator requires higher computational overheads compared to statistically simpler approaches such as the method of moments. Further, computation of the estimator under packet thinning also requires estimation, or prior knowledge, of the distribution of the flow sizes; although an empirical approximation of this can be obtained relatively easily in practice.

This article is structured as follows. We first provide some background to existing methods of network analysis, the assumed packet transit model, and our framework of analysis (SDA) in Section II. A mathematical representation of the NetFlow is presented Section III, which then allows us to define parametric NetFlow likelihoods for complete and thinned traffic. We then provide two of our key contributions, consistency and relative information loss, in Section IV. Our third contribution, the optimal family of models, is presented in Section V. Sections VI and VII are respectively dedicated to comparative analyses of the NetFlow estimator on synthetic data, and an application to real data. We finally conclude with a discussion.

II. Related Work

A. Existing Methods for Traffic Analyses

Significant progress has been made on methods for assessing network volume [7], [11]–[16] and traffic classification [8], [17], [18]. Inferential methods for analysing traffic timing, however, are less developed. Existing approaches such as series inversion [3], [7], wavelets [6], [19], empirical distributional estimation [20], cluster analysis [21], time series [22], and principal component analysis [23] typically fit models using empirical characteristics. In some cases, the intention is to simply detect deviations from typical traffic behaviour [5], [19], [24] or produce elementary network statistics [24], [25]. [6] and [26] provide simple statistical schemes whereby the parameters of a particular family of models could be identified using the method-of-moments.

Methods which are applicable to thinned network traffic, and other adversarial contexts, tend to focus on estimating network volume [11], [14], [15]. However, those which assess packet propagation are limited in their flexibility and inferential use since they typically fit secondary characteristics such as moments [3], [6], [7], [20]. Analyses of flow aggregated data often fail to account for packet thinning [2], [8], [24].

The approach we develop here accounts for both packet thinning and flow summarisation when modelling patterns of traffic propagation within the likelihood framework.

B. The Bartlett–Lewis Traffic Model

Renewal processes form a natural context for traffic analysis and have been used extensively [3], [6], [7], [20], [26]–[28]. However, [28] argues that simple renewal processes are limited in their ability to jointly model within-flow *burstiness* and interactions between flows. [6] address these limitations through the use of branching renewal processes, namely, the Bartlett–Lewis process.

The Bartlett–Lewis process is a sub-class of cluster renewal processes generated by two concurrent processes. The main and subsidiary processes respectively define the Poissonian arrival of clusters and, conditionally on the arrival of each cluster, the arrival of individual points within the cluster. The subsidiary processes are finite unidirectional random walks whose origin is the main arrival. Contextually, the first packet in each flow forms the main process, whilst the subsequent packets form translated simple finite renewal processes. Superimposing the main process and all its subsidiary processes then yields the observed traffic.

Through this framework we can develop likelihoods for both complete and flow aggregated data. Packet thinning in the context of Bartlett–Lewis processes has not been explicitly studied, but the general results of [20] can be applied quite naturally.

C. Symbolic Data Analysis

Symbolic data analysis is a relatively new field of statistics which models distributions as its fundamental datum [29]. Observations in classical statistics are typically points in a Euclidean space, having no internal variation. However, aggregation of classical data into distributional *symbols* yields objects with internal variation [30], [31]. The simplest symbol is the extremal interval, obtained by mapping a set of random variables $X_1, \ldots, X_n$ to its extrema $S = [\min X_i, \max X_i]$. The remaining $n-2$ points are then latently distributed within the interval S [30].

The methods of SDA are designed to analyse distributional forms such as random intervals and histograms [9], [30]–[34], and weighted lists [29], [31]. SDA techniques can consequently achieve computational and storage efficiency without sacrificing statistical validity. Many common statistical procedures have been extended to symbolic data including regression [31], [34]–[36], likelihood-based inference [9], [32], [33], [37], principal component analysis [38], [39], clustering [40], and time series analysis [38]. SDA has been applied to a broad range of fields including meteorology [37], finance [9], [37], medicine [9], [35], agriculture [34], ecology [41], and climatology [33].

We follow the approach of [9] which constructs the marginal likelihood for a symbol by accounting for the aggregation function applied to the underlying data and its generative model. We can establish an equivalence between symbolic and network traffic data by respectively treating packets and their NetFlows as the underlying classical data and its aggregated summaries.

Proposition 2.1 ([9]): The (non-normalised) symbolic likelihood function is

$$\mathcal{L}_S(s;\theta,\vartheta) = \int_{\mathcal{X}} f_{S|\boldsymbol{X}}(s;\boldsymbol{x},\vartheta)\, g_{\boldsymbol{X}}(\boldsymbol{x};\theta)\, \mathrm{d}\boldsymbol{x}, \tag{1}$$

where $f_{S|\boldsymbol{X}}$ is the conditional density of S given the classical data $\boldsymbol{X}$, and $g_{\boldsymbol{X}}$ is the joint density $\boldsymbol{X}$.

The likelihood $\mathcal{L}_S$ permits fitting of the underlying model $g_{\boldsymbol{X}}$ when only observing the aggregated summary S. This allows us to fit models for packet-level data when observing only the flow aggregates.

The symbolic likelihood *reduces* to the classical likelihood as the granularity of the aggregation function becomes more fine, and hence, can be seen as an approximation to the classical likelihood, where the process of summarising may induce some loss of information [9]. This property similarly holds in our definition of the NetFlow (Definitions 3.2), which is in essence an extremal interval on unidirectional data.

III. A New Estimator

A. The NetFlow Likelihood

Classical parametric likelihood-based for renewal processes uses the set of inter-renewals as observed data for inference [42]. Inference on entire *sessions* can then be made by collating the inter-renewals from all observed flows. However, this method is not immediately applicable once the flows have been aggregated as the individual inter-renewals are then lost. We address this problem by first observing NetFlows as a particular type of interval-valued random variable. We then adapt the likelihood of the renewal process for NetFlows using the likelihood in (1).

We first define an appropriate mathematical analogue for the NetFlow.

Definition 3.1: Consider a sequence of inter-renewals $\boldsymbol{X} = (X_i)_{i=1}^{M}$ of random length M embedded within the set of positive finite sequences $\mathbb{R}_+^\infty$. The NetFlow summary S is the image of the aggregation function

$$\begin{aligned}\varphi &: \mathbb{R}_+^\infty \to \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{N}\\ \varphi(\boldsymbol{X}) &= \left(X_1, \sum_{k=2}^{M} X_k, M\right)\\ &= (S_f, S_d, M)\\ &= S.\end{aligned}$$

The random elements S_f, S_d, and M respectively define the temporal distance between consecutive flows, the flow duration, and flow size. To distinguish between the complete and packet thinned settings, we denote the image of the function φ to be the *sampled* NetFlow $\widetilde{S}$ when the argument $\widetilde{\boldsymbol{X}}$ is a sequence of inter-renewals obtained from a thinned traffic, with $\widetilde{S} = (\widetilde{S}_f, \widetilde{S}_d, \widetilde{M})$. The mapping, however, remains identical in each scenario. The elements of the sampled Netflows are naturally bounded by its generating sequence so that $\widetilde{S}_d \le S_d$ and $\widetilde{M} \le M$. The set of NetFlows is then obtained by aggregating each flow.

We can now derive the NetFlow likelihood.

Proposition 3.2: Let $\boldsymbol{X} = (X_i)_{i=1}^{M}$ and $S = \varphi(\boldsymbol{X})$ respectively denote a sequence of inter-renewals and its associated NetFlow. Suppose that each inter-renewal X_i has density $g_i(\cdot;\theta_i)$, for all $i = 1, \ldots, M$. Then the NetFlow likelihood is

$$\mathcal{L}_S(S;\boldsymbol{\theta},\nu) = g_1\left(S_f;\theta_1\right)\mathcal{G}\left(S_d;\boldsymbol{\theta}'\right)p_M\left(M;\nu\right), \qquad (2)$$

where p_M is the mass function for M,

$$\mathcal{G}\left(\cdot;\boldsymbol{\theta}'\right) = g_2 * \cdots * g_M\left(\cdot;\boldsymbol{\theta}'\right),$$

$\boldsymbol{\theta} = (\theta_i)_{i=1}^{M}$, $\boldsymbol{\theta}' = (\theta_i)_{i=2}^{M}$, and $f * g$ denotes the convolution of the densities f and g.

Proof: See Appendix A. □

The NetFlow likelihood in (2) is a representative model for typical flow aggregates. The distribution of each inter-renewal is identical if the packet arrivals are defined by a simple renewal process. If, however, the packets arrive via a Bartlett–Lewis process, so that X_1 has Exponential density $f(\cdot;\lambda)$ and $X_2, \ldots, X_M$ have some common density $g(\cdot;\theta)$, then the NetFlow likelihood simplifies to

$$\mathcal{L}_S(S;\lambda,\theta;\nu) = f\left(S_f;\lambda\right)g^{*(M-1)}\left(S_d;\theta\right)p_M\left(M;\nu\right),$$

where $g^{*(M-1)}$ is the $(M-1)$-fold self-convolution of g and $\theta_1 = \lambda$.

The NetFlow likelihood for an entire session is

$$\begin{aligned}\mathcal{L}_S\left(\boldsymbol{S};\lambda,\theta,\nu\right) = \prod_{i=1}^{n} f\left(S_{f_i};\lambda\right)g^{*(M_i-1)}\left(S_{d_i};\theta\right)\\ \times p_M\left(M_i;\nu\right),\end{aligned}$$

since flows are independent, where S_i denotes i-th Netflow and $\boldsymbol{S} = (S_1, \ldots, S_n)$.

The NetFlow likelihood derived in Proposition 3.2 assumes that there is no underlying thinning. The likelihood requires some modification when traffic is thinned prior to aggregation.

Proposition 3.3: Let $\boldsymbol{X} = (X_i)_{i=1}^{M}$ be a sequence of inter-renewals and suppose that each arrival is retained with some constant known probability $q \in (0,1)$, yielding the sampled inter-renewal sequence $\tilde{\boldsymbol{X}} = (\tilde{X}_i)_{i=1}^{\tilde{M}}$. Let $\tilde{S} = \varphi(\tilde{\boldsymbol{X}})$ denote the sampled Netflow associated with $\tilde{\boldsymbol{X}}$. Let

$$p_{\tilde{M}|M}\left(\tilde{m}|m\right) = \binom{m}{\tilde{m}}q^{\tilde{m}}\left(1-q\right)^{m-\tilde{m}},$$

$$p_{J,K|M,\tilde{M}}\left(j,k|m,\tilde{m}\right) = \frac{\binom{k-j-1}{\tilde{m}-2}}{\binom{m}{\tilde{m}}},$$

$$\begin{aligned}\text{and } \mathcal{G}_{j,k}\left(\tilde{s}_f,\tilde{s}_d;\boldsymbol{\theta}\right) = g_1 * \cdots * g_j\left(\tilde{s}_f;(\theta_i)_{i=1}^{j}\right)\\ \times g_{j+1} * \cdots * g_k\left(\tilde{s}_d;(\theta_i)_{i=j+1}^{k}\right).\end{aligned}$$

The sampled NetFlow likelihood is then

$$\begin{aligned}\mathcal{L}_{\tilde{S}}\left(\tilde{S};\boldsymbol{\theta}\right) = \sum_{m,j,k} p_M\left(m;\nu\right)p_{\tilde{M}|M}\left(\tilde{M}|m\right)\\ \times p_{J,K|M,\tilde{M}}\left(j,k|m,\tilde{m}\right)\\ \times \mathcal{G}_{j,k}\left(\tilde{S}_f,\tilde{S}_d;\boldsymbol{\theta}\right),\end{aligned} \qquad (3)$$

where

$$\sum_{m,j,k} = \sum_{m \ge \tilde{M}} \sum_{j=1}^{m-\tilde{M}} \sum_{k=j+\tilde{M}-1}^{m}.$$

Proof: See Appendix B. □

The conditional mass function $p_{\tilde{M}|M}$ provides the Binomial probability of sampling $\tilde{M}$ packets from the original flow of M packets. The indices j and k respectively define the location of the first and last sampled packets with respect to their position in the original flow. The conditional mass function $p_{J,K|M,\tilde{M}}$ provides the joint probability of observing a particular pair of locations for the first and last sampled packets within the original flow. The joint conditional density for the first sampled inter-renewal $\tilde{S}_f$ and the sampled flow duration $\tilde{S}_d$ is obtained through the function $\mathcal{G}_{j,k}$. We note that the densities of $\tilde{S}_f$ and $\tilde{S}_d$ respectively require the computation of a j-fold and $(k-j)$-fold convolution of densities.

The computational cost of evaluating (3) is primarily determined by the structure of the flow sizes M and the number of packets which are sampled $\tilde{M}$ — which is itself a function of the sampling rate q. The likelihood in (3) can be computed reasonably efficiently if p_M is known *a priori* and does not give significant mass to large flow sizes. However, computational overheads will be high if large latent flow sizes are considered, and may otherwise require approximation.

As before, we can obtain the Bartlett–Lewis representation of the sampled NetFlow likelihood by setting g_1 to be the Exponential density and letting $g_k = g$ for $k \geq 2$, so that

$$\mathcal{G}_{j,k}\left(\tilde{S}_f, \tilde{S}_d; \lambda, \theta\right) = g_1 * g^{*(j-1)}\left(\tilde{S}_f; \lambda, \theta\right) \\ \times g^{*(k-j)}\left(\tilde{S}_d; \theta\right).$$

A sessional sampled NetFlow likelihood requires an average of $n!$ combinations of (3) since the ordering of the original flows cannot be determined from sampled arrivals without additional marks. However, these additional marks can often be obtained in practice through, for example, packet sequence numbers [7]. The sampled NetFlows can then be ordered according to these additional marks, thereby defining the exact sequence of flows and eliminating the need to take the aforementioned average of $n!$ combinations of flows. However, estimation in such circumstances is still somewhat complex. For example, suppose that the network only contains two flows A and B — arriving in that order — and that their respective non-empty sampled flows are $\tilde{A}$ and $\tilde{B}$. Flow-level parameters can be estimated using the sampled NetFlow likelihood in (3) if $\tilde{A}$ is observed before $\tilde{B}$. If, however, $\tilde{B}$ is observed before $\tilde{A}$, then the effective first sampled inter-renewal time for $\tilde{A}$ is actually the sum of the distance from $\tilde{B}$ to the origin and $\tilde{A}$ to $\tilde{B}$. We must hence establish the probability that $\tilde{B}$ is observed prior to $\tilde{A}$ and account for the latent convolution of packets in $\tilde{A}$ when establishing the likelihood for the sampled flow $\tilde{B}$. This sequential computation naturally increases as the number of flows increases, and the computation will be especially complex if the ordering of the sampled flows does not match the actual ordering of flows.

In practice, we may only be concerned with estimating the packet-level model, i.e. the temporal propagation of packets within each flow. It is then sufficient to only consider information obtained from the sampled flow durations $\tilde{S}_d$. We can then construct a *restricted* sampled NetFlow likelihood which is capable of evaluating the packet-level model without having to also consider the original order in which the flows arrived. Suppose that the generating renewal process is *simple*. Re-defining the sampled Netflow to be $\tilde{S} = (\tilde{S}_d, \tilde{M})$, we can then write the *restricted* NetFlow likelihood

$$\mathcal{L}_{\tilde{S}}\left(\tilde{S}; \theta, \nu\right) = \sum_{m=\tilde{M}}^{\infty} \sum_{k=\tilde{M}-1}^{m-1} p_M\left(m; \nu\right) p_{\tilde{M}|M}\left(\tilde{M}|m\right) \\ \times p_{K|M,\tilde{M}}\left(k|m, \tilde{M}\right) g^{*(k)}\left(\tilde{S}_d; \theta\right),$$

where

$$p_{K|M,\tilde{M}}\left(k|m, \tilde{M}\right) = (m-k)\frac{\binom{k-1}{\tilde{M}-2}}{\binom{m}{\tilde{M}}}$$

is the conditional probability of k inter-renewals existing between the original location of the first and last sampled packets. This implies that there were $k-1$ packets in the original flow between the first and last sampled packets, of which, $\tilde{m}-2$ and $k-\tilde{m}-3$ were respectively sampled and *not* sampled. We note that the quantity of available information is underutilised since the element $\tilde{S}_f$ is not considered in our evaluation. However, its omission may yield some benefit in reducing the required summation and simplifying computations, especially when assessing the packet-level model is of primary concern. Denoting the i-th sampled NetFlow by $\widetilde{S}_i$ and writing $\widetilde{\boldsymbol{S}} = (\widetilde{S}_1, \ldots, \widetilde{S}_n)$, the restricted sessional sampled NetFlow likelihood becomes

$$\mathcal{L}_{\widetilde{S}}\left(\widetilde{\boldsymbol{S}}; \theta\right) = \prod_{i=1}^{n} \mathcal{L}_{\widetilde{S}}\left(\widetilde{S}_i; \theta\right).$$

B. The NetFlow Estimators

Consider the log-likelihoods

$$\ell_n\left(\boldsymbol{S}; \theta\right) := \frac{1}{n}\log \mathcal{L}_S(\boldsymbol{S}; \theta) \\ \text{and } \widetilde{\ell}_n\left(\widetilde{\boldsymbol{S}}; \theta\right) := \frac{1}{n}\log \mathcal{L}_{\widetilde{S}}\left(\widetilde{\boldsymbol{S}}; \theta\right). \tag{4}$$

We define the NetFlow Maximum Likelihood Estimators (MLEs) to be the parameters $\hat{\theta}_S$ and $\hat{\theta}_{\widetilde{S}}$ which *maximise* the log-likelihoods in (4). More concretely, we write that

$$\hat{\theta}_S := \underset{\theta \in \Theta}{\arg\sup}\ \ell_n\left(\boldsymbol{S}; \theta\right) \\ \text{and } \hat{\theta}_{\widetilde{S}} := \underset{\theta \in \Theta}{\arg\sup}\ \widetilde{\ell}_n\left(\widetilde{\boldsymbol{S}}; \theta\right). \tag{5}$$

IV. Attributes of the Estimator

Suppose that the packet-level inter-renewals are obtained from the density function $g_X(\cdot; \theta_0)$. If $\boldsymbol{X} = (X_i)_{i=1}^{N}$ denotes the concatenation of all packet-level inter-renewals from every flow in the network, then the standard MLE $\hat{\theta} \in \Theta$ is the point which maximises the log-likelihood $\sum_{i=1}^{N} \log g_X(X_i; \theta)$. Consistency and efficiency are useful properties of $\hat{\theta}$ [43]. We shall show that the NetFlow estimators in (5) are also consistent, and that we may specify its degree of efficiency relative to the standard MLE.

Although specific to renewal processes, our development of consistency and relative information loss is novel since these properties of aggregated likelihoods are absent in existing SDA literature, e.g. [9], [10], [32], [33], [41].

We restrict our analysis to the packet-level in order to avoid the factorial growth in computation required to also consider flow parameters. The following results assume that the packet renewal model g_X and its sequence of self-convolutions $g_X^{*(k)}$ satisfy standard regularity conditions [43, p. 449]. They also require that the series

$$\xi(x;\theta) = \sum_{k=1}^{\infty} p_{K|M,\tilde{M}}\left(k|m,\tilde{m}\right) g_X^{*(k)}(x;\theta) \tag{6}$$

is jointly uniformly convergent in $\mathbb{R}_+ \times \Theta$ when the possible flow sizes are unbounded.

A. Consistency of the NetFlow Estimator

The following proposition extends the consistency of the standard MLE to the NetFlow MLE. Hence, parameter estimation using only aggregated NetFlows will still yield asymptotically accurate results.

Proposition 4.1: The NetFlow MLEs $\hat{\theta}_S$ and $\hat{\theta}_{\tilde{S}}$ are consistent for $\theta_0 \in \Theta^0$, where Θ^0 is the interior of the parameter space Θ.

Proof: See Appendix C. □

B. Efficiency of the NetFlow Estimator

The standard MLE is asymptotically Normally distributed with variance $(NH)^{-1}$ under some weak regularity conditions, where H and N are respectively the Fisher information of $g_X(\cdot;\theta_0)$ and the number of observed inter-renewals in the session [44]. We can adapt this result for the NetFlow MLEs by considering the Fisher information of the marginal densities of the flow durations f_{S_d} and $f_{\tilde{S}_d}$. The NetFlow MLEs will converge to θ_0 slower than the standard MLE since the estimator is computed from aggregated and thinned data. However, we can identify the number of NetFlows n which will yield a given degree of fidelity of the NetFlow estimators compared to the standard MLE.

We shall simply consider the difference in the asymptotic variances of the estimators, since they are each asymptotically Normally distributed with common mean. The determinant of the covariance matrices provides a comparable functional and is identified with its *general variance*. If the standard MLE is computed over k flows – so that $N = k\bar{M}_r$, where $\bar{M}_r$ is the mean number of packet-level inter-renewals per flow – then the MLE has covariance matrix $\Sigma = \left(k\bar{M}_r H\right)^{-1}$. Similarly, if we denote the information matrix of the flow duration by I and compute the NetFlow MLEs over n NetFlows, then we can express the covariance matrix of the NetFlow estimator by $\Upsilon = (nI)^{-1}$.

Proposition 4.2: Let g_X and f_{S_d} be the respective densities for packet-level inter-renewals and flow durations, and let their respective Fisher information be H and I. Let $\bar{M}_r$ be the average number of inter-renewals per flow computed from k flows, so that the number of packets per flow $M = M_r + 1$. Suppose that the standard MLE $\hat{\theta}$ is computed from $N = k\bar{M}_r$ inter-renewals. The number of NetFlows n which should be observed so that the efficiency of the NetFlow MLE $\hat{\theta}_S$ is within an ε relative tolerance of the standard MLE $\hat{\theta}$, for some minimum probability $1-\eta$, satisfies the inequality

$$n_+ < n < n_-,$$

where

$$n_\pm = \pm k \left(\frac{|H|/|I|}{(1\pm\varepsilon)^2}\right)^{1/d} \log\left(\frac{2}{\eta}\mathbb{E}\left[\exp\left(\pm\bar{M}_r\right)\right]\right)$$

and d is the dimension of the parameter space $\Theta \subseteq \mathbb{R}^d$.

Proof: See Appendix D. □

Proposition 4.2 can be extended to thinned traffic by instead computing the Fisher information matrix I with respect to the density of sampled flow durations $f_{\tilde{S}_d}$. We typically only care to satisfy $n > n_+$ since greater efficiency is generally desirable and additional NetFlows can be obtained freely. In such cases — where a one-sided bound suffices — the value of η in the bounds for Proposition 4.2 can be replaced by 2η. An explicit implementation of Proposition 4.2 is provided in Example 2.

V. Exact Inference With the NEF

The NetFlow likelihood requires a user specified model g_X for the inter-renewals. We wish to determine if there is a family of models for which NetFlow aggregation does not forfeit any inferential capacity. We show that the *Natural Exponential Family* (NEF) satisfies this criterion.

Definition 5.1: The Natural Exponential Family is a sub-class of the Exponential Family of distributions whose natural parameter and sufficient statistic are the identity. Respectively writing $h(x)$ and $A(\theta)$ for the base measure and log-partition function, its density can be written as

$$f_X(x;\theta) = h(x)\exp\left(\theta x - A(\theta)\right). \tag{7}$$

[45] show that the k-fold convolution of NEF has density

$$f_X^{*(k)}(x;\theta) = h_k(x)\exp\left(\theta x - kA(\theta)\right), \tag{8}$$

where h_k is the base measure for the k-fold self-convolved density $f_X^{*(k)}$. Convolutions of NEF random variables scale the sufficient statistic identically to the NetFlow aggregation, yielding the following lemma.

Lemma 5.2: Inference for the standard and NetFlow MLEs is identical when the supplied model g_X is of the Natural Exponential Family.

Proof: See Appendix E. □

Lemma 5.2 only holds for networks *without* packet thinning, although we expect that the NEF also performs comparatively better in practice than other models with thinned traffic.

This result provides a practical restriction on the optimality of model fitting since, in most cases, the sufficient statistics for distributions outside of the NEF are not derivable from the standard NetFlow summary. For example, the sufficient statistic for the shape parameter of the Gamma distribution is $\log x$. The exact sufficient statistics of convolved Gamma random variables cannot be recovered from the flow duration $S_d = \sum_{i=2}^{M+1} X_i$.

VI. Simulations

We now explore the performance of the NetFlow MLEs on various synthetic networks. [6] empirically show that packet transit can be described by the Poisson–Gamma class of the Bartlett–Lewis process, so that the temporal distance between consecutive flows and packets are respectively Exponentially and Gamma distributed. [27] provide an algorithm to generate this process. These models were also utilised by [26] and so we adopt them here for our simulated analyses. The parameters used to generate our simulated datasets are adopted from the empirical analyses of [26]. We restrict our attention to assessing packet level characteristics in order to avoid large computational overheads.

Both generation and analysis of the synthetic data in the following two examples are performed on the same machine. The machine has four Intel Core$^{\text{TM}}$ i7-6700 CPUs at 3.50 GHz and 16 GB of RAM.

Example 1: In this example, we compare the performance of the NetFlow MLE against the method-of-moments estimators of [6]. Our results show that these moments-based estimators fail to converge for the established network, unlike the NetFlow MLE. We also provide a naïve modification of the estimators of [6] which ensures convergence.

We first describe the data generating process. The empirical analysis of [26] assumes a Pareto distribution for the flow sizes with minimum $M = 1$, yielding a flow size shape parameter of $k = 1.02$. However, the Pareto distribution is, in fact, the continuous analogue for the Zeta distribution, whose discrete support reflects the underlying phenomena. We obtain the Zeta shape parameter $\kappa = 2.012085$ through reverse-engineering of the average flow size and optimisation of the Zeta mean, implying that the flow sizes have finite mean but infinite variances. We assign a finite Gamma renewal process for the packet-level process so that the distance between consecutive within-flow packets are Gamma distributed with parameters $(\alpha_0, \beta_0) = (0.6, 526.32)$.

The inter-renewals for each flow $\boldsymbol{X}$ are obtained by first sampling the Zeta distributed random flow size and then generating a sequence of independent Gamma distributed inter-renewals. The NetFlow $S = \varphi(\boldsymbol{X})$ is then computed. Realised data for an entire session is represented by $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\{s_1, \ldots, s_n\}$. Though we do not assess the flow-level model, we otherwise note that its standard and NetFlow MLE would coincide.

In [6] respectively define point estimators for α_0 and β_0 through the empirical coefficient of variation and a weighted average packet intensity. The shape estimator is explicitly defined by $\hat{\alpha} = (\overline{x}/\hat{\sigma}_x)^2$, where $\overline{x}$ and $\hat{\sigma}_x$ are respectively the empirical mean and standard deviation of all packet-level inter-renewals. Denoting the i-th flow duration by s_{d_i}, the rate estimator is then defined as $\hat{\beta} = \hat{\alpha} \sum_{i=1}^{n} w_i \varrho_i$, with weights $w_i = m_i / \sum_{j=1}^{n} m_j$, and packet intensities $\varrho_i = m_i / s_{d_i}$. While computationally fast, this approach is sequential, requires the complete set of inter-renewals, and is specific to the Gamma distribution.

We now supply equivalent moments-based estimators which use only aggregated data. Let Y_i be the mean inter-renewal of the i-th flow. We define the estimator $\breve{\alpha} = (\overline{y}/\hat{\sigma}_y)^2\, \overline{1/m}$, where $\overline{y}$ and $\hat{\sigma}_y$ are respectively the sample mean and standard deviation of the mean inter-renewal times, and $\overline{(m-1)^{-1}} = n^{-1} \sum_{i=1}^{n} 1/(m_i - 1)$ is the mean of the reciprocal flow sizes (minus 1). Substituting $\breve{\alpha}$ for $\hat{\alpha}$ into the previous rate estimator yields $\breve{\beta} = \breve{\alpha} \sum_{i=1}^{n} w_i \varrho_i$. We also compute naïve estimators $\hat{\beta}^* = \hat{\alpha}/\overline{x}$ and $\breve{\beta}^* = \breve{\alpha}/\overline{z}$, where $\overline{z} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} m_i$. Note that $\overline{x} = \overline{z}$.

TABLE I
Example 1: Mean Point Estimates (And Standard Errors) of (α, β) From $T = 10^3$ Replicate Synthetic Sessions of Size n

	n			
	10^0	10^2	10^4	10^6
Method-of-moments				
$\hat{\alpha}$	3.68	0.61	0.60	0.60
	(1.09)	$(\sim 10^{-3})$	$(\sim 10^{-4})$	$(\sim 10^{-5})$
$\hat{\beta}$	$\sim 10^3$	$\sim 10^4$	$\sim 10^5$	$\sim 10^7$
	$(\sim 10^3)$	$(\sim 10^4)$	$(\sim 10^4)$	$(\sim 10^6)$
$\hat{\beta}^*$	$\sim 10^3$	540.17	526.38	526.33
	$(\sim 10^3)$	(2.68)	(0.18)	(0.02)
NetFlow method-of-moments				
$\breve{\alpha}$	—	0.73	0.60	0.60
	—	$(\sim 10^{-3})$	$(\sim 10^{-4})$	$(\sim 10^{-5})$
$\breve{\beta}$	—	$\sim 10^4$	$\sim 10^5$	$\sim 10^7$
	—	$(\sim 10^3)$	$(\sim 10^4)$	$(\sim 10^6)$
$\breve{\beta}^*$	—	645.13	527.78	526.28
	—	(3.15)	(0.33)	(0.03)
NetFlow MLE				
$\hat{\alpha}_S$	$\sim 10^{30}$	0.65	0.60	0.60
	$(\sim 10^{30})$	$(\sim 10^{-3})$	$(\sim 10^{-4})$	$(\sim 10^{-5})$
$\hat{\beta}_S$	$\sim 10^{35}$	569.93	527.29	526.30
	$(\sim 10^{35})$	(4.48)	(0.34)	(0.04)

True values are $(\alpha_0, \beta_0) = (0.6, 526.32)$.

Table I presents the average point estimates and standard errors for each of the estimators for sessions of various size. Note that $\breve{\alpha}$ is not computable for sessions with only $n = 1$ flow since the variance of a single value is zero.

The estimates for the shape parameter α converge well for all methods. However, the moments-based rate estimators $\hat{\beta}$ and $\breve{\beta}$ notably fail to converge. This arises from estimating the rate β using the packet intensity ϱ, which has Inverse-Gamma distribution with infinite mean when its shape parameter $\alpha' \leq 1$. This occurs when the flow size $M + 1 \leq 2$, which has approximate probability 0.94 here. We can ensure that $\alpha' > 1$ by either setting the inter-renewal shape parameter $\alpha_0 > 1$ or by only recording flow sizes for which $M\alpha_0 > 1$. A discussion and examples of these conditions are provided in the Supplementary Materials.

As expected, the moments-based estimators $\hat{\alpha}$ and $\hat{\beta}^*$ computed over all the available data have considerably smaller variance than the estimates which are computed from the flow aggregates. However, we see that the NetFlow MLE is comparable to its moments-based counterparts when comparing estimators which only use aggregated data.

Table II displays the average volume of information and evaluation time used to compute each point estimate, highlighting

TABLE II
EXAMPLE 1: MEAN SESSION INFORMATION VOLUME (MEGABYTES) AND COMPUTATION TIME (MILLISECONDS) OVER VARIOUS SESSION SIZES n.

	n			
	10^0	10^2	10^4	10^6
Method-of-moments				
Size (MB)	$\sim 10^{-4}$	$\sim 10^{-3}$	1.83	100.78
Time (ms)	$\sim 10^{-1}$	$\sim 10^{-1}$	3	230
NetFlow method-of-moments				
Size (MB)	—	$\sim 10^{-3}$	0.09	9.31
Time (ms)	—	$\sim 10^{-2}$	$\sim 10^{-1}$	12
NetFlow MLE				
Size (MB)	$\sim 10^{-4}$	$\sim 10^{-3}$	0.09	9.31
Time (ms)	1	1	53	5 188

the trade-off between accuracy and computational speed. For the moment-based estimators, those based on aggregated data are typically at least one order of magnitude more efficient to compute than those based on the full flow data. This highlights the primary benefit of working with NetFlow session data. In contrast, the NetFlow MLE may be more expensive to evaluate since they may not be computed through simple arithmetic.

Example 2 (Gamma packet model with thinning): In the previous example, we generated a network with complete packet retention, which is rare in modern networks. The simple moments-based estimators and their aggregated equivalents are not coherent outside of this setting since they cannot be validly computed from thinned traffic. However, the sampled NetFlow MLE can be practically applied to both thinned and aggregated network data. We explore its performance here under various sampling rates.

As previously discussed, the computational cost of the sampled NetFlow MLE is principally determined by the cardinality of the flow sizes. Accordingly, we define a small sample space of the flow sizes $\mathcal{M} = \{10, 10^2, 10^3\}$. Flow sizes are randomly sampled from a truncated Zeta distribution with shape $\omega = 1$ such that

$$p_M(10) = \frac{6}{11},\ p_M\left(10^2\right) = \frac{3}{11},\ \text{and}\ p_M\left(10^3\right) = \frac{2}{11}.$$

We again generate flows whose packet-level arrivals follow a finite Gamma renewal process with parameters $(0.6, 526.32)$. Each packet arrival is independently recorded with probability q. The sampled NetFlows $\widetilde{s}_i$ are then computed from the thinned traffic.

In addition to a range of sampling rates q, we also compute the NetFlow MLE for various session sizes n, including $n_{\min}$, the lower bound in Proposition 4.2 with $\varepsilon = \eta = 0.1$. In plain language, $n_{\min}$ is the minimum number of flows needed for the NetFlow MLE to have efficiency within 10% of the standard MLE in at least 90% of instances. Computing $n_{\min}$ analytically is quite difficult since the information matrix I must be determined from the sampled flow densities. However, through numerical differentiation and Monte-Carlo integration, we respectively estimate that $n_{\min} = 81,\ 106,\ 551,\ 5\,065$, and $6\,507$ for the sampling rates $q = 10^{-k}$, for $k = 0, \ldots, 4$.

We also compute the standard MLE for a single flow with all of its inter-renewals. Table III presents the average (sampled) NetFlow MLE and standard errors for various session sizes and sampling rates. The rightmost column indicates the average number of seconds needed to compute the NetFlow MLE from $n_{\min}$ NetFlows.

The results show that the NetFlow MLE converges to the true parameter, regardless of the level of thinning. The results intuitively show that more NetFlows are required to achieve desired efficiency as the degree of packet thinning increases. This is also apparent in the increasing sequence of $n_{\min}$ which aims to provide a constant degree of efficiency for each sampling rate. The average time to compute the NetFlow MLE naturally increases with $n_{\min}$, but is notably sub-linear, though does not achieve $\log$-growth in this example. We see that point estimation is relatively time effective, requiring on average less than two minutes for a heavily thinned network with sampling rate $q = 10^{-4}$.

VII. REAL DATA ANALYSIS

We now explore the performance of the NetFlow MLE on real network data. We obtain a `pcap` file from [46] which captures one minute of network activity, observing 36 197 062 packets distributed amongst 1 811 255 flows, of which, we identify 779 788 non-trivial flows.

Timestamps are recorded at nanosecond granularity. In some cases, inter-renewals will be recorded as 0 since consecutive packets may arrive within this threshold. We assign these zero-valued inter-renewal times to be 10^{-7} seconds, on the scale of the smallest positive inter-renewal time $\delta \approx 2.38 \times 10^{-7}$ seconds.

The `pcap` trace was processed using a high performance cluster to extract the arrival times, source and destination addresses, and IP tag. Each job utilised 20 GB of RAM on four cores. The flows were then constructed from the aforementioned information using 10 parallel jobs, each of which utilised four cores with 8 GB of RAM each. The generation of NetFlows and parameter estimation was performed on a local machine with four Intel Core$^{\text{TM}}$ i7-6700 CPUs at 3.50 GHz and 16 GB of RAM.

A. Full Packet Retention

We first provide an analysis assuming complete data. The standard MLE requires approximately 2 hours to compute using the Gamma model and its fitted survival function is presented in Fig. 1(a). It is clear from Fig. 1 that the Gamma model is misspecified. We instead naïvely compute the sample mean and standard deviation of the log-transformed inter-renewals

$$\left(\overline{\log(x)},\ \hat{\sigma}_{\log(x)}\right) = (-8.0987,\ 4.5046)\,, \tag{9}$$

where x denotes the packet inter-renewals. These quantities correspond to the MLE for the Log-Normal distribution. Fig. 1(a) shows that the Log-Normal model with parameters (9) provides a sufficiently good fit for the packet inter-renewals.

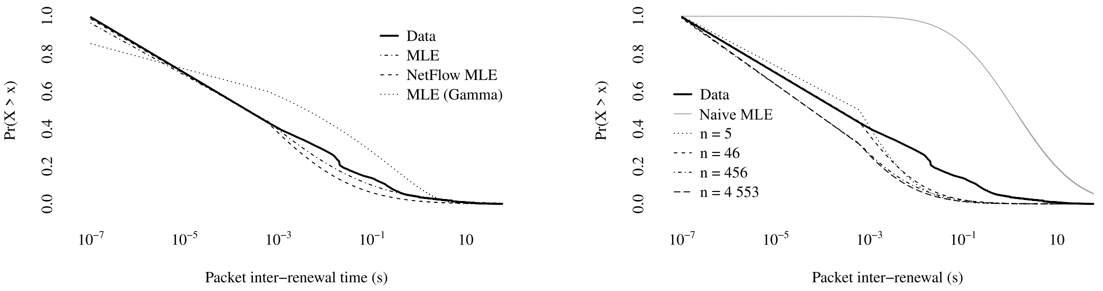
TABLE III
EXAMPLE 2: MEAN (SAMPLED) NETFLOW MLE (AND STANDARD ERRORS) OF (α, β) FOR A SYNTHETIC NETWORK WITH A PACKET-LEVEL GAMMA PROCESS

			n			
	10^0	10^1	10^2	10^3	$n_{\min}$	Time (s)
MLE $(q = 1)$						
$\hat{\alpha}$	0.72	—	—	—	—	—
	(0.01)	—	—	—	—	—
$\hat{\beta}$	703.74	—	—	—	—	—
	(14.69)	—	—	—	—	—
NetFlow MLE						
$q = 1$						
$\hat{\alpha}_S$	$\sim 10^{29}$	0.84	0.62	0.60	0.63	10^{-3}
	$(\sim 10^{28})$	(0.02)	$(\sim 10^{-3})$	$(\sim 10^{-4})$	$(\sim 10^{-3})$	
$\hat{\beta}_S$	$\sim 10^{32}$	737.37	544.33	528.88	551.38	
	$(\sim 10^{31})$	(14.72)	(2.45)	(0.76)	(3.07)	
$q = 10^{-1}$						
$\tilde{\alpha}_S$	$\sim 10^{29}$	29.22	0.64	0.60	0.63	5
	$(\sim 10^{28})$	(12.13)	$(\sim 10^{-3})$	$(\sim 10^{-3})$	$(\sim 10^{-3})$	
$\tilde{\beta}_S$	$\sim 10^{33}$	$\sim 10^{4}$	561.20	529.01	553.17	
	$(\sim 10^{32})$	$(\sim 10^{4})$	(3.89)	(1.09)	(3.89)	
$q = 10^{-2}$						
$\tilde{\alpha}_S$	$\sim 10^{28}$	$\sim 10^{3}$	5.49	0.63	0.66	20
	$(\sim 10^{27})$	(175.05)	(0.88)	$(\sim 10^{-3})$	$(\sim 10^{-3})$	
$\tilde{\beta}_S$	$\sim 10^{32}$	$\sim 10^{6}$	$\sim 10^{3}$	552.61	575.42	
	$(\sim 10^{32})$	$(\sim 10^{5})$	(765.00)	(3.68)	(6.24)	
$q = 10^{-3}$						
$\tilde{\alpha}_S$	$\sim 10^{28}$	$\sim 10^{3}$	188.22	2.59	0.66	97
	$(\sim 10^{26})$	(143.03)	(10.73)	(0.23)	$(\sim 10^{-3})$	
$\tilde{\beta}_S$	$\sim 10^{31}$	$\sim 10^{6}$	$\sim 10^{5}$	$\sim 10^{3}$	578.06	
	$(\sim 10^{30})$	$(\sim 10^{5})$	$(\sim 10^{3})$	(195.88)	(5.97)	
$q = 10^{-4}$						
$\tilde{\alpha}_S$	$\sim 10^{28}$	$\sim 10^{3}$	160.17	2.77	0.65	115
	$(\sim 10^{26})$	(120.78)	(8.97)	(0.26)	$(\sim 10^{-3})$	
$\tilde{\beta}_S$	$\sim 10^{31}$	$\sim 10^{6}$	$\sim 10^{5}$	$\sim 10^{3}$	573.49	
	$(\sim 10^{30})$	$(\sim 10^{5})$	$(\sim 10^{3})$	(224.11)	(5.33)	

Estimates are obtained over a range of session sizes n and packet thinning rates q. presented values are obtained from $T = 10^3$ replicate datasets with non-trivial flows. true values are $(\alpha_0, \beta_0) = (0.6, 526.32)$. The right-most column shows the average time (seconds) to compute the (sampled) netflow MLE using $n_{\min} = 81,\ 106,\ 551,\ 5\,065$, and 6 507 netflows for respective sampling rates q as Shown.

The NetFlow MLE for the Log-Normal model is not immediately accessible since there are no simple, closed-form convolutions of Log-Normal random variables. We instead estimate the convolution through the *Fenton–Wilkinson* approximation [47]. Unfortunately, this approximation cannot be readily substituted into the NetFlow likelihood since the session contains several mouse flows whose durations are too small to satisfy the tail approximation. We remedy this by further aggregating the set of NetFlows into a single *session NetFlow*, obtained by taking the element-wise sum of all NetFlows. This *two-step* approximation yields the NetFlow MLE $(\hat{\mu}_S, \hat{\sigma}_S) = (-7.9511, 3.6684)$.

Fig. 1(a) shows that the NetFlow MLE slightly underestimates large scale inter-renewals, but is an otherwise satisfactory representation of the observed data. Table IV (top two rows) presents the size of each dataset and times for computing the standard and NetFlow MLE. In this instance, computation for the MLE is



(a) $q = 1$

(b) $q = 0.001$

Fig. 1. Survival functions for the observed inter-renewals and various fitted models. Fig. 1(a) supplies the survival function under complete packet retention, i.e. $q = 1$. Fig. 1(b) depicts the fitted models when each packet in the network is retained through a Bernoulli trial with probability $q = 0.001$. The thick black line indicates the survival function of the observed data and is identical in both (a) and (b). The dotted line in (a) is obtained by fitting the MLE for a Gamma model to the complete set of inter-renewals. All other supplied survival functions are fitted to the Log-Normal distribution. The solid grey line in (b) indicates the standard MLE which is naïvely applied to the sampled inter-renewals. The dotted and dashed lines in (b) correspond to the fitted models from the sampled NetFlow MLE obtained from the experiment for the given number of datapoints n.

TABLE IV
Mean Point Estimates (And Standard Errors) for the Real Data Application in Section VII

	Data volume		Data Parameters		Times (s)
	p	n	μ	σ	
$q = 1$					
NetFlow MLE	1	779 788	-8.05 —	3.70 —	6
Standard MLE	1	34 385 807	-8.10 —	4.50 —	3
$q = 0.001$					
NetFlow MLE	0.001	5	-7.41 (0.64)	2.24 (0.41)	410
	0.01	46	-7.84 (0.46)	2.73 (0.29)	714
	0.1	456	-8.90 (0.03)	3.20 (0.07)	2 718
	1	4 553	-8.79 —	2.92 —	13 938
Naïve MLE	1	19 511	0.22 —	2.42 —	0

The quantity p determines the proportion of total available information used to compute the point estimate. The number n denotes the number of datapoints which correspond to p and respectively refers to the number of inter-renewals and netflows over which the standard and netflow MLE are computed. The right-most column presents the average time (in seconds) required to compute the point estimate. Estimates with $p < 1$ are averaged over $T = 10$ resamples of the data set.

trivially faster since the point estimates can be expressed through simple arithmetic. However, we note that the file size of the complete set of inter-renewals required to compute the standard MLE is 26.1 times larger than the set of NetFlows, and that this larger dataset is typically not recorded in practice.

B. A Packet Thinned Scenario

We consider a more realistic setting by analysing packet-thinned samples of the full dataset, with sampling probability $q = 10^{-3}$. The sampled NetFlows are then computed using Definition 3.1. We observe 36 189 sampled arrivals which are distributed amongst 4 553 sampled flows.

We obtain an initial point estimate by computing the sample mean and standard deviation of the sampled inter-renewals, which we term to be the *naïve* MLE. In order to compute the sampled NetFlow MLE, we need to supply the mass function for the original flow size p_M. We compute an empirically derived truncation of the mass function to limit computational overload. We assume that the original flow sizes are restricted to $M = \lceil j \times 10^k \rceil$, where $j = 1, 2.5, 5$, and $k = 0, \dots, 5$, rounding each observed flow size to the nearest restricted flow size, and then deriving the approximate mass function from the proportion of flows rounded to each restricted flow size value. We then compute the sampled NetFlow MLE using the estimated flow size mass function and the Fenton–Wilkinson approximation for the convolution of the density of the Log-Normal distribution. However, in contrast to the scenario with full packet retention, the sampled NetFlows were not aggregated into a larger sessional sampled NetFlow. Hence, the Fenton–Wilkinson approximation is poor for sampled NetFlows with smaller sampled flow durations and sampled flow sizes. The under-estimation arising from the second degree approximation is apparent in Fig. 1(b) and the value of the point estimates presented in Table IV. This contrasts the accurate convergence of the NetFlow MLE presented in Example 2 which utilised an exact convolution for the Gamma model.

The results are presented in the middle rows of Table IV and in Fig. 1(b). The quantity p indicates the proportion of the total available data used to compute the point estimate. The quantity n corresponds to p and shows the resulting number of datapoints used in the optimisation.

From Fig. 1(b) it is clear that naïvely fitting the Log-Normal distribution to the sampled inter-renewals (solid-grey line) fails to adequately describe packet transit. The sampled NetFlow MLE performs well, even with the approximations involved in computing convolutions and the mass of flow sizes. The model fit is naturally poorer than when using the full dataset, although the full dataset is typically unavailable. More accurate fits can be obtained by using finer approximations of p_M.

VIII. Discussion

We have introduced a novel method for parametric, likelihood-based inference of network packet data when the utilised data are (possibly) thinned and aggregated. The ability to jointly handle packet thinning and NetFlow aggregation within the likelihood framework, and all its inferential benefits, is a great advantage over existing methods for analysing flow data, which can only account for one of these processes. The maximum likelihood estimators themselves are consistent (with increasing numbers of NetFlows), and we have derived bounds on the number of (thinned) NetFlows needed to attain a given degree of accuracy. As a result, the NetFlow likelihoods offer a practical and flexible tool for inference on very large session datasets.

The potentially large computational cost is the price to pay for this framework. Indeed, computing the NetFlow likelihoods requires a convolution of densities which can be particularly complex when the model is not closed under convolutions. However, we have shown that, even on real data, closed-form approximations can be utilised without severe consequences. Optimising a likelihood is slower than methods that rely on simple arithmetic computation (such as the moments-based estimators of [6] and [26]). Computation also increases with higher degrees of packet thinning (i.e. low q), or with a high frequency of elephant flows, since the NetFlow likelihood function needs to consider (and integrate out) all possible latent flows which could have produced the observed, thinned flow.

Despite these limitations, the NetFlow likelihood estimators provide an effective method for network analysis when the data we have to work with is less than ideal: both heavily summarised, and heavily sub-sampled.

Appendix

A. Proof to Proposition 3.2

Proof: The joint conditional density of the first inter-renewal S_f and flow duration S_d is

$$f_{S_f,S_d|\boldsymbol{X},M}\left(s_f,s_d|\boldsymbol{x},m\right) = \delta_{s_f}\left(x_1\right)\delta_{s_d}\left(\sum_{k=2}^{m} x_k\right), \tag{10}$$

where δ_a is the Dirac measure centred at a. The sequence of inter-renewals $\boldsymbol{X}$ has conditional density

$$g_{\boldsymbol{X}|M}\left(\boldsymbol{x}|m;\boldsymbol{\theta}\right) = \prod_{i=1}^{m} g_i\left(x_i;\theta_i\right). \tag{11}$$

Substituting (10) and (11) into (1) and computing the integral over the space $\mathbb{R}_+^m$ yields the conditional likelihood

$$\mathcal{L}_{S_f,S_d|M}\left(s_f,s_d|m;\boldsymbol{\theta}\right) = g_1\left(s_f;\theta_1\right)\mathcal{G}\left(s_d;\boldsymbol{\theta}'\right). \tag{12}$$

Multiplying (12) by the mass function $p_M(m;\nu)$ then yields the result. □

B. Proof to Proposition 3.3

Proof: The conditional density for the first sampled-inter-renewal $\tilde{S}_f$ and sampled flow duration $\tilde{S}_d$ is

$$f_{\tilde{S}_f,\tilde{S}_d|\tilde{\boldsymbol{X}},\tilde{M},M}\left(\tilde{s}_f,\tilde{s}_d|\tilde{\boldsymbol{x}},\tilde{m}\right) = \delta_{\widetilde{s_f}}\left(\tilde{x}_1\right)\delta_{\tilde{s}_d}\left(\sum_{i=2}^{\tilde{m}}\tilde{x}_i\right). \tag{13}$$

In a manner similar to Appendix A, we now wish to define the conditional density of the vector of sampled inter-renewals $f_{\tilde{\boldsymbol{X}}|\tilde{M}}$. There are $\binom{m}{\tilde{m}}$ possible ways to construct a vector of sampled inter-renewals of length $\tilde{m}$ from the original vector of inter-renewals $\boldsymbol{x}$ of length m, each of which are equally likely to be obtained. Hence, we have that

$$f_{\tilde{\boldsymbol{X}}|\boldsymbol{X},M,\tilde{M}}\left(\tilde{\boldsymbol{x}}|\boldsymbol{x},m,\tilde{m}\right) = \binom{m}{\tilde{m}}^{-1}. \tag{14}$$

Noting that the sequence of inter-renewals $\boldsymbol{X}$ is independent of the number of sampled packets $\tilde{M}$, we have that the conditional density $f_{\boldsymbol{X}|\tilde{M},M} = g_{\boldsymbol{X}|M}$, as defined in (11). Substituting (11) and (14) into (1) and computing the integral over the space

$$\left\{n \in \mathbb{N} : n \leq \binom{m}{\tilde{m}}\right\} \times \mathbb{R}_+^m$$

with respect to the product of the counting and m-dimensional Lebesgue measure yields the conditional likelihood

$$\mathcal{L}_{\tilde{S}_f,\tilde{S}_d|M,\tilde{M}}\left(\tilde{s}_f,\tilde{s}_d|m,\tilde{m};\boldsymbol{\theta}\right) = \int_{\mathbb{R}_+^m} g_{\boldsymbol{X}|M}\left(\boldsymbol{x}|m;\boldsymbol{\theta}\right)\delta_{\widetilde{s_f}}\left(\tilde{x}_1\right)\delta_{\tilde{s}_d}\left(\sum_{i=2}^{\tilde{m}}\tilde{x}_i\right)\mathrm{d}\boldsymbol{x}. \tag{15}$$

Suppose that the first sampled packet corresponds with the J-th packet in the original sequence, and that the last sampled packet — the $\tilde{m}-$th sampled packet — corresponds with the K-th packet in the original sequence, where $\tilde{m} \geq 2$ and $1 \leq J < K \leq m$. It follows that $\tilde{s}_f = \sum_{i=1}^{J} x_i$ and $\tilde{s}_d = \sum_{i=J+1}^{K} x_i$. The quantities J and K are, in fact, random variables. Write $\mathbb{R}_+^m = \mathbb{R}_+^J \times \mathbb{R}_+^{K-J} \times \mathbb{R}_+^{m-K}$. Conditioning the likelihood in (15) with respect to J and K and computing the integral then yields

$$\mathcal{L}_{\tilde{S}_f,\tilde{S}_d|M,\tilde{M},J,K}\left(\tilde{s}_f,\tilde{s}_d|m,\tilde{m},j,k;\boldsymbol{\theta}\right) = \mathcal{G}_{j,k}\left(\tilde{s};\boldsymbol{\theta}\right). \tag{16}$$

Conditional on M and $\tilde{M}$, J takes values in the set

$$\Omega_J = \{j \in \mathbb{N} : j \leq M - \tilde{M} + 1\}$$

and has mass function

$$p_{J|M,\tilde{M}}(j|m,\tilde{m}) = \frac{\binom{m-j}{\tilde{m}-1}}{\binom{m}{\tilde{m}}} \tag{17}$$

since there are $\binom{m-j}{\tilde{m}-1}$ equally possible sampled flows which can be obtained when the first sampled packet corresponds to the j-th packet in the original flow. Similarly, conditional on M, $\tilde{M}$, and J, K takes values in the set

$$\Omega_K = \{k \in \mathbb{N} : J + \tilde{M} - 1 \leq k \leq M\}$$

and has mass function

$$p_{K|J,M,\tilde{M}}\left(k|j,m,\tilde{m}\right) = \frac{\binom{k-j-1}{\tilde{m}-2}}{\binom{m-j}{\tilde{m}-1}} \tag{18}$$

since there are $\binom{k-j-1}{\tilde{m}-2}$ uniformly possible which can be obtained when the first and last sampled packets respectively correspond to the j-th and k-th packets in the original flow. Integrating the product of (16), (17), and (18) over the space $\Omega_J \times \Omega_K$ with respect to the product counting measure yields that

$$\mathcal{L}_{\tilde{S}_f,\tilde{S}_d|M,\tilde{M}}\left(\tilde{s}_f, \tilde{s}_d|m, \tilde{m}; \boldsymbol{\theta}\right) = \sum_{j=1}^{m-\tilde{m}+1} \sum_{k=j+\tilde{m}-1}^{m} \frac{\binom{k-j-1}{\tilde{m}-2}}{\binom{m}{\tilde{m}}} \mathcal{G}_{j,k}\left(\tilde{s}_f, \tilde{s}_d; \boldsymbol{\theta}\right). \tag{19}$$

Taking the product of (19), p_M, and $p_{\tilde{M}|M}$ and integrating over the space

$$\{m \in \mathbb{N} : m \geq \tilde{m}\} \tag{20}$$

with respect to the counting measure then yields the result. □

C. Proof to Proposition 4.1

Proof: We first show that $\hat{\theta}_S$ is consistent. We then define conditions which give consistency for $\hat{\theta}_{\tilde{S}}$ when applying an identical approach. Firstly, we assume that Θ is locally compact and we denote a general compact neighbourhood by $\overline{\Theta}$. Let $\Lambda(s;\theta) := \mathbb{E}_{\theta_0}[\ell_1(s;\theta)]$.

Lemma 8.1: The sequence $\{\ell_n(\boldsymbol{s};\theta)\}_{n\geq 1}$ converges to $\Lambda(s;\theta)$ pointwise by the law of large numbers.

Lemma 8.2: Suppose that g_X is identifiable in Θ, i.e. $g_X(\cdot,\theta_1) = g_X(\cdot,\theta_2) \Rightarrow \theta_1 = \theta_2$. Then $\Lambda(s,\theta)$ is uniquely maximised at θ_0.

Proof: An application of Jensen's inequality yields that

$$\mathbb{E}_{\theta_0}\left[g_X^{*(m)}(x;\theta_0)\right] > \mathbb{E}_{\theta_0}\left[g_X^{*(m)}(x;\theta)\right]$$

for all $\theta \neq \theta_0$. Hence, $\Lambda(s;\theta)$ is maximised at θ_0. This point is also unique as a consequence of identifiability. □

Definition 8.3 (Stochastic equicontinuity [48]): A sequence of real-valued random functions $\{f_n(x;\theta)\}_{n\geq 1}$ is *stochastically equicontinuous* in θ if, for all positive ε and η, there exists a positive δ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{\vartheta\in\overline{\Theta}} \sup_{|\theta-\vartheta|<\delta} |f_n(x;\vartheta) - f_n(x;\theta)| > \varepsilon\right) < \eta.$$

Lemma 8.4: The sequences $\{\ell_n(\boldsymbol{s};\theta)\}_{n=1}^{\infty}$ and $\{\Lambda(s;\theta)\}_{n=1}^{\infty}$ are stochastically equicontinuous.

Proof: The sequence $\{\Lambda(s;\theta)\}_{n\geq 1}$ is stochastically equicontinuous since $\Lambda(s;\theta)$ is constant in n and uniformly continuous in $\overline{\Theta}$.

We now prove that the sequence $\{\ell_n(\boldsymbol{s};\theta)\}_{n\geq 1}$ is stochastically equicontinuous. Firstly, let

$$\gamma(s_i;\vartheta,\theta) = \log\left(g^{*(m_i)}(s_{d_i};\vartheta)\right) - \log\left(g^{*(m_i)}(s_{d_i};\theta)\right).$$

For some positive ε and δ, we have that

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{\vartheta\in\overline{\Theta}} \sup_{\theta\in B(\vartheta,\delta)} |\ell_n(\boldsymbol{s};\vartheta) - \ell_n(\boldsymbol{s};\theta)| > \varepsilon\right) \leq \frac{1}{\varepsilon}\mathbb{E}\left[\sup_{\vartheta\in\overline{\Theta}} \sup_{\theta\in B(\vartheta,\delta)} |\gamma(s_j;\vartheta,\theta)|\right], \tag{21}$$

where

$$j = \arg\max_{i=1,\ldots,n} \mathbb{E}\left[\sup_{\vartheta\in\overline{\Theta}} \sup_{\theta\in B(\vartheta,\delta)} |\gamma(s_i;\vartheta,\theta)|\right].$$

For some positive η, we can choose $\delta' = \delta(\varepsilon,\eta)$ such that $\sup_{|\theta-\vartheta|<\delta'} |\gamma(s_j;\theta,\vartheta)| < \eta\varepsilon$. It follows that $\{\ell_n(\boldsymbol{s};\theta)\}_{n\geq 1}$ is stochastically equicontinuous since (21) can be bounded from above by η. □

An application of Theorem 2.1 of [48] shows that ℓ_n converges to Λ uniformly in probability, and hence $\hat{\theta}_S$ is consistent.

The proof for the sampled NetFlow estimator $\hat{\theta}_{\tilde{S}}$ is identical if the cardinality of the flow sizes is bounded. However, we require that the series (6) is jointly uniformly convergent when the cardinality is countably infinite. We can then pursue the same method of proof with the stated condition. □

D. Proof to Proposition 4.2

Proof: Every covariance matrix induces a hyper-ellipsoid whose semi-axes lengths are equal to its eigenvalues [49]. The volume of the induced hyper-ellipsoid is hence proportional to the square root of the determinant of the generating covariance matrix. Hyper-ellipsoids with smaller volumes then imply that the generating random vector has smaller *variance*. Hence, we may geometrically compare the efficiency of the NetFlow and standard estimators by observing the volume of their associated hyper-ellipsoids. For a d-dimensional parameter space $\Theta \subseteq \mathbb{R}^d$, we have that

$$\begin{aligned} |\Sigma| = \left|\left(k\bar{M}_r H\right)^{-1}\right| &= \left(k\bar{M}_r\right)^{-d} |H|^{-1} \\ \text{and } |\Upsilon| = \left|(nI)^{-1}\right| &= n^{-d}|I|^{-1}. \end{aligned}$$

We wish to determine the number of NetFlows n which we should observe so that for some positive ε we have that

$$\left|\frac{\operatorname{Vol}\hat{\theta}_S}{\operatorname{Vol}\hat{\theta}} - 1\right| \equiv \left|\frac{|\Upsilon|^{1/2}}{|\Sigma|^{1/2}} - 1\right| < \varepsilon, \tag{22}$$

where $\operatorname{Vol}\boldsymbol{X}$ denotes the volume of the hyper-ellipsoid induced by the random vector $\boldsymbol{X}$. The ratio of the volumes of the induced hyper-ellipsoids hence satisfies the bounds

$$1 - \varepsilon < \frac{|\Upsilon|^{1/2}}{|\Sigma|^{1/2}} < 1 + \varepsilon.$$

We however note that (22) describes a random event with respect to the random variable $\bar{M}_r$. We can instead consider bounding (22) in probability such that for any positive ε and η, we have that

$$\mathbb{P}\left(\left|\frac{|\Upsilon|^{1/2}}{|\Sigma|^{1/2}} - 1\right| < \varepsilon\right) \geq 1 - \eta,$$

or equivalently,

$$\mathbb{P}\left(\left|\frac{|\Upsilon|^{1/2}}{|\Sigma|^{1/2}} - 1\right| \geq \varepsilon\right) < \eta. \tag{23}$$

This ensures that the efficiency of the NetFlow and standard MLEs are within an ε tolerance of each other for a minimum specified long run average $1 - \eta$.

To obtain the given bounds, consider the equation

$$\left|\frac{|\Upsilon|^{1/2}}{|\Sigma|^{1/2}} - 1\right| \geq \varepsilon \tag{24}$$

and let $r = n/k$ and $R = |I|/|H|$. Then, squaring both sides of (24), substituting for Σ, Υ, r, and R, and rearranging yields the equation

$$r^{-d}R^{-1}\bar{M}_r^d - 2r^{-d/2}R^{-1/2}\bar{M}_r^{d/2} + 1 - \varepsilon^2 \geq 0. \tag{25}$$

Equation (25) is a positive quadratic with respect to $\bar{M}_r^{d/2}$ and has roots

$$x_\pm = r^{d/2}R^{1/2}(1 \pm \varepsilon).$$

As a set, (24) can be rewritten as the union of disjoint sets

$$\left\{\bar{M}_r \geq x_+^{2/d}\right\} \cup \left\{\bar{M}_r \leq x_-^{2/d}\right\},$$

so that (23) can be replaced by

$$\mathbb{P}\left(\bar{M}_r \geq x_+^{2/d}\right) + \mathbb{P}\left(\bar{M}_r \leq x_-^{2/d}\right) < \eta.$$

Applying Chernoff's bound to the first summand, bounding the limit by $\eta/2$, and rearranging for r yields

$$r > \left(\frac{R^{-1}}{(1+\varepsilon)^2}\right)^{1/d} \log\left(\frac{2}{\eta}\mathbb{E}\left[\mathrm{e}^{\bar{M}_r}\right]\right).$$

Substituting for r and R and rearranging for n then yields the lower bound

$$n > k\left(\frac{|H|/|I|}{(1+\varepsilon)^2}\right)^{1/d} \log\left(\frac{2}{\eta}\mathbb{E}\left[\mathrm{e}^{\bar{M}_r}\right]\right) =: n_+.$$

Performing a similar set of operations to the latter summand yields the upper bound

$$n < -k\left(\frac{|H|/|I|}{(1-\varepsilon)^2}\right)^{1/d} \log\left(\frac{2}{\eta}\mathbb{E}\left[\mathrm{e}^{-\bar{M}_r}\right]\right) =: n_-.$$

Note that the two-sided interval is only satisfied if

$$\mathbb{E}\left[\mathrm{e}^{\bar{M}_r}\right]\mathbb{E}\left[\mathrm{e}^{-\bar{M}_r}\right] > \left(\frac{\eta}{2}\sqrt[d]{\frac{1+\varepsilon}{1-\varepsilon}}\right)^2.$$

□

E. Proof to Proposition 5.2

Proof: Recalling the densities in (7) and (8), for a sequence of inter-renewals $\boldsymbol{x}$ and its associated NetFlow $s = (s_f, s_d, m+1)$, we have the likelihoods

$$\mathcal{L}(\boldsymbol{x};\theta) \propto \exp\left(\theta\sum_{i=1}^{m+1} x_i - (m+1)A(\theta)\right) \text{ and}$$

$$\mathcal{L}_S(s;\theta) \propto \exp\left(\theta\left(s_f + s_d\right) - (m+1)A(\theta)\right),$$

which are identical since $\sum_{k=1}^{m+1} x_k = s_f + s_d$. □

Acknowledgment

This research includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney.

References

[1] Cisco, "Cisco visual networking index: Forecast and trends, 2017–2022," San Jose, California, USA, White Paper, Feb. 2019.

[2] R. Hofstede *et al.*, "Flow monitoring explained: From packet capture to data analysis with netflow and IPFIX," *IEEE Surv. Tut.*, vol. 16, no. 4, pp. 2037–2064, 2014.

[3] N. Hohn and D. Veitch, "Inverting sampled traffic," *IEEE/ACM Trans. Netw.*, vol. 14, no. 1, pp. 68–80, Feb. 2006.

[4] R. Jurga and M. Hulbój, "Packet sampling for network modelling," CERN, Geneva, Switzerland, Tech. Rep., CH-1211, Dec. 2007.

[5] L. Bin, L. Chuang, Q. Jian, H. Jianping, and P. Ungsunan, "A netflow based flow analysis and monitoring system in enterprise networks," *Comput. Netw.*, vol. 52, no. 5, pp. 1074–1092, 2008.

[6] N. Hohn, D. Veitch, and P. Arby, "Cluster processes: A natural language for network traffic," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2229–2244, Aug. 2003.

[7] N. Antunes and V. Pipiras, "Estimation of flow distributions from sampled traffic," *ACM Trans. Modelling Perform. Eval. Comput. Syst.*, vol. 1, no. 3, pp. 1–28, 2016.

[8] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, and J. Solé-Pareta, "Analysis of the impact of sampling on netflow traffic classification," *Comput. Netw.*, vol. 55, no. 5, pp. 1083–1099, 2011.

[9] B. Beranger, H. Lin, and S. Sisson, "New Models for Symbolic Data Analysis," 2018, *arXiv:1809.03659*.

[10] X. Zhang, B. Beranger, and S. Sisson, "Constructing likelihood functions for interval-valued random variables," *Scand. J. Statist.*, vol. 47, no. 1, pp. 1–35, 2020.

[11] N. Duffield, C. Lund, and M. Thorup, "Estimating flow distributions from sampled flow statistics," *IEEE/ACM Trans. Netw.*, vol. 13, no. 5, pp. 933–946, Oct. 2005.

[12] D. Veitch and P. Tune, "Optimal skampling for the flow size distribution," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3075–3099, Jun. 2015.

[13] L. Yang and G. Michailidis, "Sampled based estimation of network traffic flow characteristics," in *Proc. IEEE INFOCOM 2007-26th IEEE Int. Conf. Comput. Commun.*, 2007, pp. 1775–1783.

[14] Y. Chabchoub, C. Fricker, F. Guillemin, and P. Robert, "On the statistical characterisation of flows in internet traffic with application to sampling," *Comput. Commun.*, vol. 33, no. 1, pp. 103–112, Jan. 2010.

[15] B. Ribeiro, D. Towsley, T. Ye, and J. Bolot, "Fisher information of sampled packets: An application to flow size estimation," in *Proc. 6th ACM SIGCOMM Conf., Internet Meas.*, 2006, pp. 15–26.

[16] N. Brownlee and K. Claffy, "Understanding internet traffic streams: Dragonflies and tortoises," *IEEE Commun. Mag.*, vol. 40, no. 10, pp. 110–117, Oct. 2002.

[17] S. S. Kundu, K. Pal Basu, and S. Das, "Fast classification and estimation of internet traffic flows," in *Passive and Active Network Measurement* (Lecture Notes in Computer Science Series), S. K. Uhlig Papagiannaki and O. Bonaventure, Eds., vol. 4427. Berlin, Germany: Springer, 2007, pp. 155–164.

[18] Y. Miao, Z. Ruan, L. Pan, J. Zhang, Y. Xiang, and Y. Wang, "Comprehensive analysis of network traffic data," in *Proc. IEEE Int. Conf. Comput. Inf. Technol.*, 2016, pp. 423–430.

[19] S. Stoev, M. Taqqu, C. Park, and J. S. Marron, "On the wavelet spectrum diagnostic for hurst parameter estimation in the analysis of internet traffic," *Comput. Netw.*, vol. 48, no. 3, pp. 423–445, 2005.

[20] N. Antunes and V. Pipiras, "Probabilistic sampling of finite renewal processes," *Bernoulli*, vol. 17, no. 4, pp. 1285–1326, 2011.

[21] J. Kim, A. Sim, B. Tierney, S. Suh, and I. Kim, "Multivariate network traffic analysis using clustered patterns," *Computing*, vol. 101, no. 4, pp. 339–361, Apr. 2019.

[22] C. You and K. Chandra, "Time series models for internet data traffic," in *Proc. 24th Conf. Local Comput. Netw*. Lowell, MA, USA, 1999, pp. 164–171.

[23] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 32, no. 1, pp. 61–72, Mar. 2004.

[24] A. Proto, L. Alexandre, M. Batista, I. Oliveira, and A. Cansian, "Statistical model applied to netflow for network intrusion," in *Transactions on Computational Science XI: Special Issue on Security in Computing*, M. L. Gavrilova, C. J.K. Tan, and E. D. Moreno, Eds., Berlin Heidelberg, Germany: Springer-Verlag, 2010, vol. 2, pp. 179–191.

[25] Y. Lee, W. Kang, and H. Son, "An internet traffic analysis method with mapreduce," in *Proc. IEEE/IFIP Netw. operations Manage. Symp. Workshops*, 2010, pp. 357–361.

[26] N. Antunes, V. Pipiras, P. Abry, and D. Veitch, "Small and large scale behavior of moments of poisson cluster processes," *ESAIM: Probability Statist.*, vol. 21, pp. 369–393, 2017.

[27] B. González-Arévalo and J. Roy, "Simulating a poisson cluster process for internet traffic packet arrivals," *Comput. Commun.*, vol. 33, no. 5, pp. 612–618, Mar. 2010.

[28] C. Williamson, "Internet traffic measurement," *IEEE Internet Comput.*, vol. 5, no. 6, pp. 70–74, Nov./Dec. 2001.

[29] L. Billard and E. Diday, "Symbolic data analysis: Definitions and examples," Univ. Georgia, Athens, USA, White Paper, 2004.

[30] L. Billard and E. Diday, "From the statistics of data to the statistics of knowledge: Symbolic data analysis," *J. Amer. Stat. Assoc.*, vol. 98, no. 462, pp. 470–487, 2003.

[31] L. Billard and E. Diday, *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Southern Gate, Chichester, West Sussex, U.K.: Wiley, 2007.

[32] J. Le-Rademacher and L. Billard, "Likelihood functions and some maximum likelihood estimators for symbolic data," *J. Stat. Plan. Inference*, vol. 141, no. 4, pp. 1593–1602, Apr. 2011.

[33] T. Whitaker, B. Beranger, and S. Sisson, "Composite likelihood methods for histogram-valued random variables," *Statist. Comput.*, vol. 30, pp. 1459–1477, 2020.

[34] T. Whitaker, B. Beranger, and S. Sisson, "Logistic regression models for aggregated data," *J. Comput. Graphical Statist.*, vol. 30, no. 4, pp. 1049–1067, 2021, doi: 10.1080/10618600.2021.1895816.

[35] L. Billard and E. Diday, "Symbolic regression analysis," in *Classification, Clustering, and Data Analysis*, K. Jajuga, A. SokoŁ owski, and H.-H. Bock, Eds., Berlin, Heidelberg: Springer, Jul. 2002, pp. 281–288.

[36] E. Neto and F. de Carvalho, "Centre and range method for fitting a linear regression model to symbolic interval data," *Comput. Statist. Data Anal.*, vol. 52, no. 3, pp. 1500–1515, Jan. 2008.

[37] P. Brito and A. Silva, "Modelling interval data with normal and skew-normal distributions," *J. Appl. Statist.*, vol. 39, no. 1, pp. 3–20, Jan. 2012.

[38] M. Noirhomme-Fraiture and P. Brito, "Far beyond the classical data models: Symbolic data analysis," *Stat. Anal. Data Mining*, vol. 4, no. 2, pp. 157–170, Mar. 2011.

[39] V. Lauro and F. Palumbo, "Principal component analysis of interval data: A symbolic data analysis approach," *Comput. Statist.*, vol. 15, no. 1, pp. 73–87, Sep. 2000.

[40] R. Verde, "Clustering methods in symbolic data analysis," in *Classification, Clustering, and Data Mining Applications*, D. F. R. Banks McMorris, P. Arabie, and W. Gaul, Eds., Chicago, USA: Springer, Jul. 2004, pp. 299–317.

[41] H. Lin, M. Caley, and S. Sisson, "Estimating global species richness using symbolic data meta-analysis," *Ecography,* vol. 2022, no. 3 Mar. 2022, Art. no. e05617.

[42] A. Karr, *Point Processes and Their Statistical Inference*, 2nd ed. New York, NY, USA: Marcel Dekker Inc.4, 1991.

[43] E. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York, NY, USA: Springer-Verlag, 1998.

[44] R. Höpfner, *Asymptotic Statistics : With a View to Stochastic Processes*, 1st ed. Berlin, Germany: Walter de Gruyter GmbH & Co KG, 2014.

[45] C. Morris and K. Lock, "Unifying the named natural exponential families and their relatives," *Amer. Statistician*, vol. 63, no. 3, pp. 247–253, Aug. 2009.

[46] CAIDA, "The CAIDA UCSD anonymized internet traces <20190117-1315500>," 2019. Accessed: Apr. 22, 2020. [Online]. Available: http://www.caida.org/data/passive/passive_dataset.xml

[47] N. Marlow, "A normal limit theorem for power sums of independent random variables," *Bell Syst. Tech. J.*, vol. 46, no. 9, pp. 2081–2089, Nov. 1967.

[48] W. Newey, "Uniform convergence in probability and stochastic equicontinuity," *Econometrica*, vol. 59, no. 4, pp. 1161–1167, 1991.

[49] Y. Tong, *The Multivariate Normal Distribution* (Ser. Springer Series in Statistics Series), 1st ed. New York, NY, USA: Springer-Verlag, 1990.