K-Sparse Approximation for Traffic Histogram Dimensionality Reduction

Atef Abdelkefi*, Yuming Jiang*, Xenofontas Dimitropoulos†
* Norwegian University of Science and Technology (NTNU), Norway
† Swiss Federal Institute of Technology (ETH), Zurich

***Abstract*—Traffic histograms play a crucial role in various network management applications such as network traffic anomaly detection. However, traffic histogram-based analysis suffers from the *curse of dimensionality*. To tackle this problem, we propose a novel approach called K-*sparse approximation*. This approach can drastically reduce the dimensionality of a histogram, while keeping the approximation error low. K-sparse approximation *reorders* the traffic histogram and uses the top-K coefficients of the reordered histogram to approximate the original histogram. We find that after reordering the widely-used histograms of source port and destination port exhibit a *power-law* distribution, based on which we establish a relationship between K and the resulting approximation error. Using a set of traces collected from a European network and a regional network, we evaluate our K-*sparse approximation* and compare it with a well-known entropy-based approach. We find that the power-law property holds for different traces and time intervals. In addition, our results show that K-*sparse approximation* has a unique property that is lacking in the entropy-based approach. Specifically, K-*sparse approximation* explicitly exposes a tradeoff between compression level and approximation accuracy, enabling to easily select a desired settlement point between the two objectives.**

I. Introduction

Traffic histograms portray the number of packets, bytes or flows observed during a time interval for different values of a traffic feature, such as source IP address or port number. They play a crucial role in various network management applications as they provide different views of traffic characteristics useful for traffic accounting [7], traffic anomaly detection [10], and resource provisioning [11] among other applications.

However, traffic histogram-based analysis suffers from the *curse of dimensionality*. This problem is evident in the following example. In the European network of GEANT2 [1], a record of 15-minute traffic data includes approximately 10^9 flows distributed over 2^{16} ports and 2^{32} IP addresses [14]. These are formidable numbers for histogram-based analysis as they result in long vectors, which are hard to process. It is worth highlighting that while traffic sampling is widely used to support online traffic monitoring, it does not significantly reduce the number of entries in a histogram.

To address the dimensionality problem, it is crucial to develop dimensionality reduction techniques for traffic histograms. Surprisingly, the literature has little work focusing on dimensionality reduction of traffic histograms. The several existing approaches, e.g. [9], [10], [12], [15], use histogram dimensionality reduction techniques that are tailored to specific applications. Except for [12] where a histogram is simply summarized into a single entropy value, the other approaches reduce the dimensionality of a traffic histogram based on heuristics that typically involve an empirically-selected threshold value. Specifically, the threshold value is a relative uncertainty value in [15], a traffic volume value in [9], and a port number, e.g., port 1024, in [10]. However, while working well under the suggested threshold values for their respective purposes, these approaches do not provide explicit relationship between the selected threshold value and the approximation error. Since the threshold value affects the number of selected components and hence the approximation error, which is called the "information-loss" tradeoff in this paper, lacking a relationship for the "information-loss" tradeoff makes the decision of using a certain threshold value highly empirical and limits the broader applicability of traffic histogram analysis.

In this paper we propose a novel approach for dimensionality reduction of traffic histograms called K-*sparse approximation*. The proposed approach is based on an explicit relationship between the number of chosen coefficients and the approximation error. Specifically, K-*sparse approximation reorders* a traffic histogram and uses the top-K coefficients to approximate it. Our technique has its root on the finding that after reordering, the traffic histograms of popular features exhibit a *power-law* distribution. Based on this, we establish a relationship between the approximation error and K. Compared to a state-of-the-art technique based on entropy [15], our technique has the key advantage that it allows to easily control the trade-off between the desired compression level and the affordable approximation error. Using traffic traces from several locations in a European network and from a regional network, the proposed K-*sparse approximation* approach is investigated and compared with the related entropy-based approach. In addition, we investigate the impact of sampling under both approaches. The results are promising showing that K-*sparse approximation* is useful for tackling the curse of dimensionality in traffic histogram analysis.

The main contributions of the paper are the following:

- We propose K-*sparse approximation*, a novel histogram dimensionality reduction technique that uses a (possibly

978-3-901882-48-7

very) small number of coefficients K to accurately approximate large histograms.
- We observe that the histograms of port numbers, when sorted, decay according to a power law. Based on this observation, we derive a relationship between K and the approximation error.
- We compare K-*sparse approximation* with a state-of-the-art entropy-based approach and show that K-*sparse approximation* results in a more stable K.
- We illustrate that K-sparse approximation works effectively both with sampled and un-sampled traffic.

The remainder is organized as follows. Sec. II introduces basic concepts and the datasets used in the investigation in Sec. III. Sec. IV presents the entropy-based approach. We introduce our proposed approach in Sec. V. We expose results in In Sec. VI, discuss related work in Sec. VII and conclude in Sec. VIII.

II. Traffic Histogram and Approximation Error

A traffic histogram is the distribution of the traffic volume (in terms of flows, packets or bytes) over all possible values or *coefficients* of a feature. A feature is a field in the packet header, such as source port, or a *function* of some header field values, such as the destination IP prefix. While there are many features that may be analyzed, we focus here on source and destination port numbers.

Specifically, consider an observation window T and a feature X, which has n possible distinct values in the range $[1, 2, \dots n]$. Assume that the amount of traffic corresponding to the i-th feature value is x_i, where $i = 1, \dots, n$. The traffic histogram of X for this observation window is $(x_1, x_2, ...x_n)$.

With a bit of abuse of notation, we shall also use X to represent both a feature and a corresponding traffic histogram $(x_1, x_2, ...x_n)$, although the correct interpretation should be clear given the context. X_K denotes the traffic histograms that results from setting to zero all the initial n feature values, but for K selected values. Using X_K to approximate the original traffic histogram X, the *approximation error* (σ_K) is defined as:

$$\sigma_K = \frac{||X - X_K||_2}{||X||_2} \tag{1}$$

where $||Y||_2$ denotes the Euclidean norm of Y.

As it is clear from Eq. 1, the approximation error takes values between 0 and 1. In general, the larger the subset X_K, the lower the approximation error. Hence, there is a trade-off between K and σ_K, or in other words, a trade-off between the information kept (decided by K) and the information lost (reflected by σ_K). A "good" dimensionality reduction approach should keep a balance between the parameters involved in the "information-loss" trade-off, and particularly, should select a subset X_K subject to a desired approximation error.

This triggers the key research question of which K coefficients to choose among the set of all coefficients of a traffic histogram.

To answer this question, we reorder the coefficients of a histogram such that in the reordered histogram, denoted by $X' \equiv (x'_{(1)}, ..., x'_{(n)})$, the coefficients are in the non-increasing order, i.e. $x'_{(1)} \geq x'_{(2)} \geq \cdots \geq x'_{(n)}$. With this, the following result helps select K coefficients.

Proposition 1: In approximating X, the top K coefficients of the reordered histogram X' give the least approximation error (σ_K) among all possible choices of K coefficients.

Proof: Based on Eq. 1, for a given K, larger coefficients result in smaller approximation error σ_K.

Formally, let $X'_K \equiv (x'_{(1)}, ..., x'_{(K)}, 0...0)$ where $x'_{(1)}, ..., x'_{(K)}$ are the top K coefficients of X'. Note that for the corresponding K coefficients in X, the approximation error is equivalently computed using the equation:

$$\sigma_K = \frac{||X' - X'_K||_2}{||X'||_2} \tag{2}$$

where $||X'||_2 = ||X||_2$ is easily verified.

For ease of expression, in the following we consider a case where only one element in X'_K is replaced, and the general case can be extended from this. Specifically, consider another set of K coefficients, denoted by $X''_K \equiv (x'_{(1)}, \dots, x''_{(i)}, \dots, x'_{(K)}, 0...0)$. In X''_K, $K-1$ components are the same as in X'_K. Without loss of generality, suppose the only different element $x''_{(i)}$ is the jth $(j > K)$ coefficient in X'_K. Since the K coefficients in X'_K are the top coefficients, we must have $x'_{(i)} \geq x''_{(i)} = x'_{(j)}$. Then, with simple manipulation, we have:

$$\begin{aligned} & ||X' - X''_K||_2 - ||X' - X'_K||_2 \\ = \; & [(x'_{(i)} - x'_{(j)})^2 + {x'_{(i)}}^2] - {x'_{(j)}}^2 \\ = \; & 2{x'_{(i)}}^2 - 2x'_{(i)}x'_{(j)} \geq 0 \end{aligned}$$

with which the result follows. ■

The above result lays the foundation for the dimensionality reduction approaches discussed in the next sections.

III. The Dataset and Histograms of Interest

The datasets used in this paper are traffic traces collected from the GEANT2 network [1]. GEANT2 is a pan-European backbone network that interconnects and provides Internet access to European NRENs (National Research and Educational Networks). The traffic traces were collected from four network links: (1) a peering link between the Internet and the Frankfurt router in GEANT2 (Trace $\mathcal{A}$); (2) a peering link between the Internet and the Vienna router in GEANT2 (Trace $\mathcal{B}$); (3) a peering link between the Internet and the Amsterdam router in GEANT2 (Trace $\mathcal{C}$); and (4) a peering link between the Internet and the Copenhagen router in GEANT2 (Trace $\mathcal{D}$).

The four traces were collected during a 33-day measurement period in June – July 2011, and involve traffic flow records recorded over 15-minute measurement time bins at a sampling rate of $1/1000$.

In this paper, we are interested in histograms of source (srcport) and destination (dstport) port numbers over one-day (24h) intervals. Fig. 1 displays an example histogram of the number of flows over source ports for a 24h interval in the measurement period. Port numbers larger than 1024 are clipped for visibility.

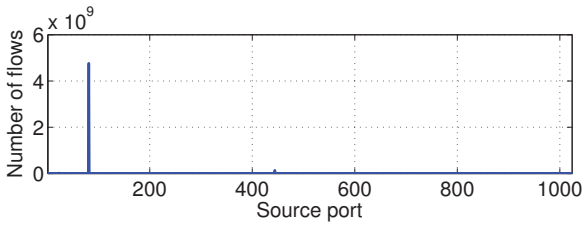


Fig. 1. Example traffic histogram

IV. The Entropy Based Approach

In this section, we describe the entropy based approach initially introduced in [15] to derive X_K. Specifically, the entropy based approach helps select K based on which we compute X'_K from the reordered traffic histogram. This approach relies on and makes use of two concepts: *sample entropy* and *relative uncertainty (RU)*.

Consider a random variable X that may take n discrete values. Suppose we observe X for S times. Let x_i denote the number of observed times when X takes the value i, for $i = 1, \ldots, n$. Then, the sample entropy of X, denoted by $H(X)$, is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)), \tag{3}$$

and its relative uncertainty defined as:

$$RU(X) = \frac{H(X)}{H_{max}(X)} \tag{4}$$

where $p(x_i) = \frac{x_i}{S}$, and $H_{max}(X) \equiv \log \min(n, S)$. Here, H_{max} denotes the maximum entropy of X, since $2^{H_{max}(X)}$ represents the maximum possible number of unique values that X may take.

Putting into the context of a traffic histogram, X represents the considered feature (e.g. source port or destination port), n denotes the maximum number of possible values of the feature, e.g. 2^{16} source / destination ports in IPv4, x_i the number of observed flows in a certain time interval which have source / destination port number i, $i = 1, \ldots, n$; and S the total number of observed flows in this time interval.

An important property of RU is the following (see also [15] and references therein). In the case $S \leq n$, let A denote the subset of observed values in X, in which, $p(x_i) > 0$ for $i \in A$. Then, it is known that $RU(X) = 1$ if and only if $|A| = S$ and $p(x_i) = \frac{1}{S}$. In the case $S > n$, then $RU(X) = 1$ if and only if $x_i = \frac{S}{n}$, which implies $p(x_i) = \frac{1}{n}$. Both cases tell that if the relative uncertainty calculated from a set of observed values x_i is close to one, i.e. $RU(X) \approx 1$, the random variable X approximately has a uniform distribution over the distinct feature values corresponding to the (non-zero subset of the) observed x_i.

In other words, $RU(X) = 1$ implies no variation among (the non-zero subset of) x_i. On the other hand, if few of the observed values x_i have higher probability to appear, then the distribution of X becomes skewed and $RU(X) < 1$. In the extreme case when only one observed x_i has none-zero value, $RU(X) = 0$.

The novel idea of the entropy based approach has its root on the above property of RU. Particularly, it repeatedly calculates the relative uncertainty (RU) of $(x'_{(K+1)}, x'_{(K+2)}, ..., x'_{(n)}) \equiv R(K)$, starting from $K = 1$. The iteration stops when $RU(R)$ is close to 1. In other words, the entropy method focuses on and evaluates if the remaining tail is close to uniform. It repeats this procedure by increasing K in each step and stops when the tail is almost uniform. Fig. 2, illustrates an example of RU in function of the K value during one of the measurement days. It shows that RU increases as K increases, until it becomes relatively stable and close to 1 when K is high, i.e, the tail of the reordered histogram is close to be uniformly distributed.

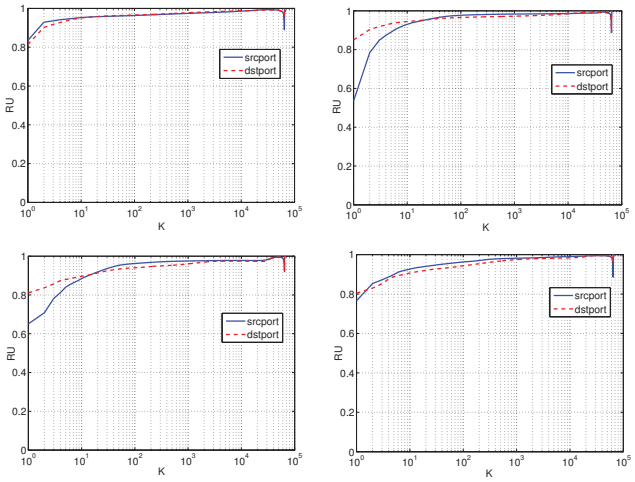


Fig. 2. RU as a function of the K value

As suggested in [15], we choose $RU(R) > 0.9$ as the stopping threshold, with which the resulting K value indicates the number of coefficients $x'_{(1)}, x'_{(2)}, ..., x'_{(K)}$ used to approximate the original traffic histogram.

A pseudo-code of this method can be found in [15] and is reproduced in Algorithm 1, where β is a "cut-off" threshold used in [15] to decide if the remaining set R is close to uniformly distributed.

Fig. 3 illustrates the resulting number of K by applying the entropy based approach on the daily traffic histograms over the 33 measurement days for the four collected traces. The figure depicts that while the entropy based approach is effective in reducing the dimension of the traffic histograms, the number of coefficients found using this approach is highly variable. For example, the number of K coefficients can dramatically vary from several, to tens or even hundreds between successive days within the same trace. The number also varies between traces, as well as between features.

It is worth highlighting that the entropy based approach does not have a targeted σ_K. The high variation in the obtained K values implies that they may lead to significant differences in the approximation error using the obtained top

Algorithm 1 Entropy-based approach

Input: *Reordered* traffic distribution histogram $X' = (x'_{(1)}, x'_{(2)}, ...x'_{(n)})$

Output: The value of K

$S = \emptyset, R = X', i = 0, \beta = 2\%$
compute cond. prob. dist. of R and its $RU(R)$
while $RU(R) \leq 0.9$ **do**
  $\beta = \beta \times 2^{-i}; i++$
  for $x'_{(j)} \in R$ **do**
    if $p(x'_{(j)}) \geq \beta$ **then**
      $S = S \bigcup \{x'_{(j)}\}; R = R - \{x'_{(j)}\};$
    end if
  end for
  compute cond. prob. dist. of R and its $RU(R)$
end while
$K = card(S)$

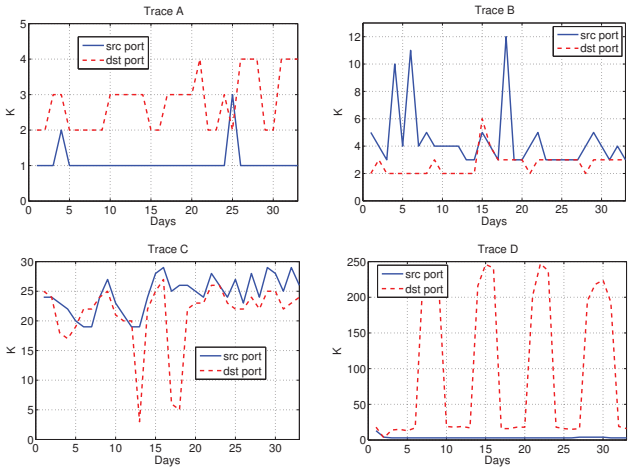


Fig. 3. The K value derived from the entropy-based approach

K coefficients. This motivates us to propose a new approach for dimensionality reduction in the next section.

V. K-Sparse Approximation

Recall the discussion in Sec. II on approximation error. This section presents the proposed approach that provides an explicit link between the chosen K coefficients and the corresponding approximation error. Based on this, given a required σ_K, the top K coefficients are readily found.

A. Power-Law Traffic Distribution

The key idea of the proposed approach is to explore the characteristics of the reordered histogram $X' \equiv (x'_{(1)}, x'_{(2)}, ..., x'_{(n)})$, $x'_{(1)} \geq x'_{(2)} \geq \cdots \geq x'_{(n)}$. For this purpose, Fig. 4 illustrates the traffic histograms of source and destination ports over a 24h period from the collected traces. To produce these figures (and similar ones in Sec. VI), traffic histograms have been normalized (rescaled to unit norm) and reordered such that, the histograms coefficients are sorted in order of decaying magnitude.

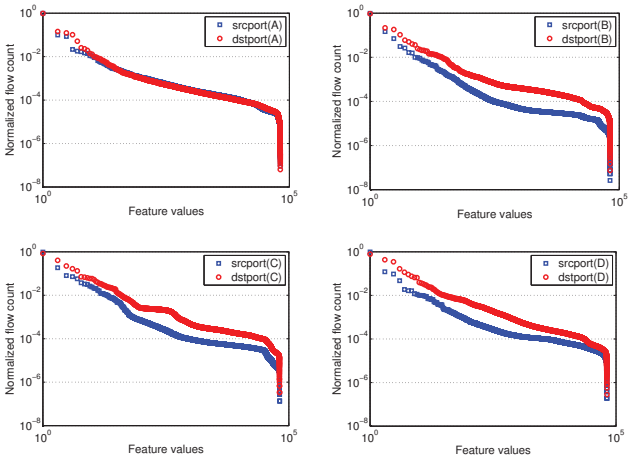


Fig. 4. Sorted traffic histograms over a 24h-period for the 4 traces

It is worth re-pointing to Fig. 1 and comparing with it. Recall that Fig. 1 is an original traffic histogram that displays the number of flows per feature value, arranged in the order of the feature values. While Fig. 1 does indicate that the traffic is only concentrated on some source ports, Fig. 4 additionally illustrates that a *power-law* model is a plausible fit to the reordered histograms, due to the linear tendency in the log-log scale.

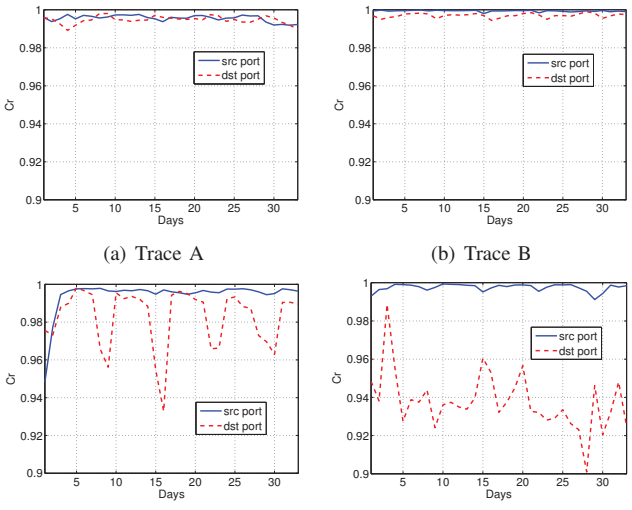
To get a better insight into the validity of our observation, we conducted a linear regression to fit the reordered traffic histograms in the log-log scale. The fitting technique is based on the least-square error method [2]. The validity of the approximation is indicated by the correlation coefficient (Cr) which measures the quality of a least-square fit to the original data. Cr is a number between -1 and 1. A Cr value equal to 1 implies a perfect linear correlation between the original data point and the fitting data point, while a Cr value of 0 indicates no correlation between the original data point and the fit.

Fig. 5 illustrates the correlation coefficients for each of the 24hs traffic histograms over the entire measurement interval. The figure depicts that unlike Trace $\mathcal{A}$ and $\mathcal{B}$, where the Cr exceeds 0.99 for both source and destination ports histograms, the correlation coefficients experienced for destination ports in trace $\mathcal{C}$ and $\mathcal{D}$ are relatively lower and more fluctuating. However, the power law fit holds for all instances of traffic histograms during the entire measurement interval, since the figure shows that the correlation coefficients are relatively high (> 0.9) for all collected traces.

B. Approximation with Top K-Coefficients

Assuming that a reordered traffic histogram is drawn from a distribution that follows a *power law*, one can extract a few top K-coefficients to approximate the traffic histogram. We present in this section the proposed methodology.

Let $X \equiv (x_1, x_2, \ldots, x_n)$ be the traffic histogram and its coefficients. For ease of expression in the following, we suppose that these coefficients are *normalized* (against the Euclidean norm of X) and rescaled to unit norm. In addition,



(a) Trace A (b) Trace B

(c) Trace C (d) Trace D

Fig. 5. Correlation coefficients over the measurement interval

let $X' \equiv (x'_{(1)}, x'_{(2)}, \ldots, x'_{(n)})$ represent the reordered histogram where the coefficients are *sorted* in order of decaying magnitude ($x'_{(1)} \geq x'_{(2)} \geq ...x'_{(n)}$).

The observation from Fig. 4 indicates that $x'_{(i)}$, $(i = 1, 2, \ldots)$, decay according a power law:

$$x'_{(i)} \leq G \cdot i^{-\alpha} \tag{5}$$

where G is a normalization constant and α is a scaling parameter with $\alpha \geq 1$. This observation lays the foundation for the following analysis.

A histogram X' with a power law decay can be approximated by the first few K coefficients i.e. the top-K coefficients, by keeping the largest K coefficients and setting the remaining to zero. The resulting histogram, denoted by X'_K, is a compressed representation of the original histograms X' and X. We call X'_K the top K **sparse approximation**. Thanks to the rapid decay of its coefficients, in our measurement traces, we typically have $K \ll N$ under a small approximation error.

Particularly, it can be verified that the top K-sparse approximation has an approximation error [5]:

$$\sigma_K = ||X - X_K||_2 = ||X' - X'_K||_2 \leq (\frac{s}{\alpha})^{(-1/2)} G K^{(-s)} \tag{6}$$

where $s = \alpha - \frac{1}{2}$, and the normalization constant G for the power-law distribution has the following expression:

$$G = \alpha - 1. \tag{7}$$

The scaling parameter α can be estimated using the *maximum likelihood estimator* (MLE) and is given by the following expression [8]:

$$\alpha = 1 + n \left[\sum_{i=1}^{n} \ln x'_{(i)} \right]^{-1}. \tag{8}$$

Finally, given the scaling parameter α estimated with Eq. 8 and the normalization parameter G calculated with Eq. 7, the following result is readily obtained with Eq. 6, which establishes an explicit relationship between the required approximation error σ_K and the corresponding K number of top coefficients:

$$K \geq [\frac{\sigma_K}{(\frac{s}{\alpha})^{(-1/2)} G}]^{(-\frac{1}{s})} \tag{9}$$

with which, the traffic histogram is approximated by $(x'_{(1)}, \ldots x'_{(K)})$.

Summarizing the above discussion, we present a pseudo-code of the proposed approach in Algorithm 2. We highlight that while Eq. 9 provides an *explicit* link between K and the approximation error, the entropy-based approach does not. This link maintains a tuning knob for the control of the "information-loss" tradeoff, which is a key advantage of the proposed approach.

Algorithm 2 K-sparse approximation approach

Input: *Reordered normalized* traffic histogram: $X' = (x'_{(1)}, x'_{(2)}, \ldots, x'_{(n)})$; required approximation error σ_K

Output: The value of K.

$\alpha = 1 + n \left[\sum_{i=1}^{n} \ln x'_{(i)} \right]^{-1}$

$G = \alpha - 1$

$K \geq [\frac{\sigma_K}{(\frac{s}{\alpha})^{(-1/2)} G}]^{(-\frac{1}{s})}; s = \alpha - \frac{1}{2}$

C. Stability of the Scaling Parameter

The scaling parameter α controls the approximation / compression performance as indicated by Eq. 9 and Eq. 7. A higher (lower) scaling parameter increases (decreases) the compressibility of a histogram. In the following, we give insights into the characteristics of this parameter.

Fig. 6 illustrates the scaling parameter's cumulative distribution functions for reordered 24-hour histograms of source and destination ports over the entire measurement interval and for all collected traces. The figure indicates that the scaling parameter varies for different 24-hour time intervals, features, and datasets. However, the variation is small.

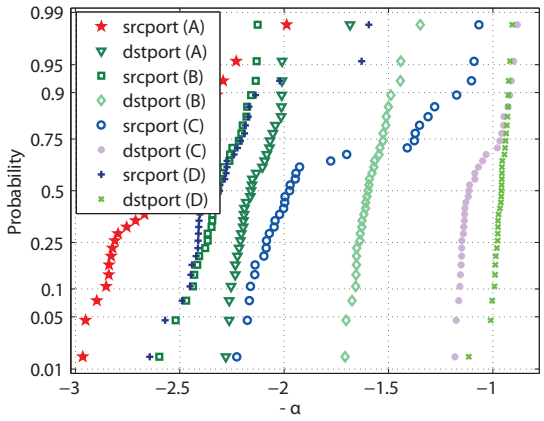


Fig. 6. Scaling parameter distributions

The mean and standard deviation of the estimated scaling parameters are presented in Table I. Interestingly, the table shows that the estimated scaling parameters exhibit relative stability over time with low standard deviation. In addition, Table I indicates that the traffic distribution over source ports experiences for all traces faster decay with larger scaling parameter than that for destination ports. This means that traffic is concentrated on a larger number of top destination ports rather than source ports. Moreover, the table shows that the collected traces exhibit different compressibility levels. The number of flows per destination port has the slowest decay, i.e, the smallest scaling parameter, for trace $\mathcal{D}$ and the highest scaling parameter for trace $\mathcal{A}$. In contrast, the variation of the scaling parameter for histograms of source ports is much smaller.

TABLE I
ESTIMATED MEAN AND STANDARD DEVIATION FOR THE SCALING PARAMETERS FOR THE FOUR TRACES

Feature	Trace $\mathcal{A}$	Trace $\mathcal{B}$	Trace $\mathcal{C}$	Trace $\mathcal{D}$
SrcPort	2.61 ± 0.22	2.13 ± 0.11	1.79 ± 0.38	2.28 ± 0.21
DstPort	2.13 ± 0.11	1.58 ± 0.08	1.06 ± 0.09	0.96 ± 0.03

To further investigate the impact of the scaling parameter on the resulting K value, we illustrate in Table II the mean and the standard deviation (STD) of the K values resulting from both entropy and our sparse approximation with a targeted approximation error of 20% ($\sigma_K = 0.2$). The Table shows that, except for trace $\mathcal{C}$, the K values derived from our method in all traces experience a lower standard deviation in comparison to those derived from the entropy-based method. For example the relative standard deviation (RSD) for the K values derived from entropy in trace $\mathcal{D}$ is in the order of 105% while it is just in the order of 21% with our method. Due to the high experienced STD, we investigate, in Fig. 7, the behavior of the K values derived from our approach ($\sigma_K = 0.2$) for destination port traffic histograms per 24h over the whole measurement period in trace $\mathcal{C}$. The figure clearly illustrates a daily pattern where the resulting K value generally decreases during week days while it increases during the weekend.

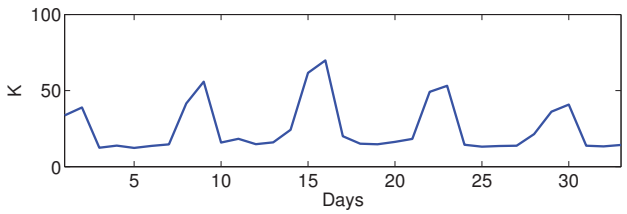


Fig. 7. Daily pattern for the K value in trace $\mathcal{C}$ for dst port

It is worth highlighting that Fig. 6, Table I and Table II all show that the scaling parameter and as a consequence the resulting K value are relatively stable over time. This is exciting and has a significant implication in practice: One may use the scaling parameter estimated in the past to pre-decide how many top K coefficients to keep in approximating the current histogram.

VI. UNDERSTANDING K-SPARSE APPROXIMATION

In this section we investigate the performance of K-sparse approximation and compare it with the entropy-based approach. The focused performance aspect is the trade-off between the selected K and the resulting approximation error.

A. K-Sparse Approximation v.s. Entropy-Based

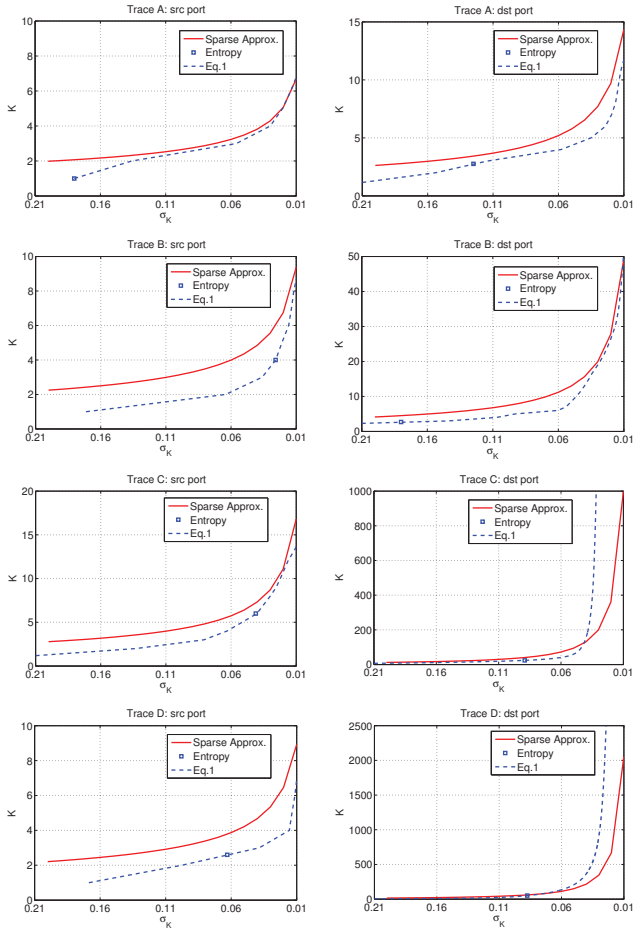


Fig. 8. K value as a function of the approximation error (GEANT2)

In Fig. 8, we plot the curves of the number K decided from K-sparse approximation in function of the approximation error (σ_K) for the four traces and for both source and destination port histograms. To obtain these curves, we vary σ_K in the interval $[0.21, 0.01]$ and calculate for each σ_K the required value of K based on Eq. 9 for each 24h traffic histogram over the measurement period. Since K varies due to small variations in the scaling parameter, as discussed in Sec. V-C, we plot in Fig. 8 the mean of obtained K in function of σ_K.

To get an overview of how well K-sparse approximation performs, we additionally draw in Fig. 8 the curves of K as a function of the approximation error directly calculated using Eq. 1, for the four traces and for both source and destination port histograms. These curves are drawn by varying the K

TABLE II
MEAN AND STANDARD DEVIATION OF K FOR ENTROPY AND SPARSE APPROXIMATION USING FOUR DIFFERENT TRACES

Approach	Feature	Trace $\mathcal{A}$	Trace $\mathcal{B}$	Trace $\mathcal{C}$	Trace $\mathcal{D}$
Entropy	Srcport	1.09 ± 0.38	4.39 ± 2.25	24.48 ± 3.06	3.45 ± 1.75
	Dstport	2.78 ± 0.78	2.69 ± 0.8	21.12 ± 5.75	97.66 ± 103.45
Sparse approximation ($\sigma_K = 0.2$)	Srcport	2.26 ± 0.23	2.58 ± 0.15	2.87 ± 3.6	2.68 ± 0.54
	Dstport	2.87 ± 0.29	4.97 ± 0.7	25.41 ± 16.62	35.76 ± 7.62

value in the obtained range while calculating for each K the corresponding σ_K using Eq. 1 for each 24h traffic histogram over the whole measurement period. One has to note that the actual approximation error varies within each of the 24-h traffic histograms; for this reason, we plot in Fig. 8 the mean of the obtained σ_K. We refer these curves the ideal tradeoff curves.

To compare with the entropy-based approach, we also draw in Fig. 8 the average value of K decided with Algorithm 1 for an RU threshold of 0.9. Since the entropy-based approach does not provide an explicit relationship between the number of coefficients and σ_K, we calculate the actual approximation error using Eq. 1 directly.

Fig. 8 illustrates that while the required number K of top coefficients increases when the targeted approximation error increases, it generally remains very small for reasonably low approximation error. For example, with only $K = 7$ top coefficients, which corresponds to only 0.01% of all coefficients, we can achieve an approximation error of 1% for source port traffic histograms in trace $\mathcal{A}$, while for destination ports in trace $\mathcal{D}$, which exhibit the worst compressibility among all traces, 16 coefficients result in an error of approximately 20%.

Fig. 8 additionally shows that the curve of K as a function of the approximation error obtained using our approach closely matches with the curve of the ideal trade-off calculated using Eq. (1). In general, the proposed approach gives a bit conservative number of coefficients for a reasonable approximation error e.g. 5%, but the difference is small, e.g. only one coefficient difference for all source port traffic histograms. Note that, a bit conservative (yet very close) estimation of K is what is desired, ensuring the targeted estimation error.

However, if the targeted approximation error is too small, e.g. smaller than 5% for destination ports in traces $\mathcal{C}$ and $\mathcal{D}$, the proposed approach may experience performance degradation. In particular, in such a case, K sparse approximation may give a too optimistic number of coefficients in comparison to the number actually required to achieve the ideal tradeoff. For example, to achieve an approximation error of 3% for destination port histograms in trace $\mathcal{C}$, 668 top coefficients are calculated from the proposed approach, while the calculation directly from Eq. 1 gives more than a thousand coefficients to achieve the same approximation error. This is caused by an inaccurate power-law fit. Indeed, in all traces as shown in Fig. 4, the lower-tail of the distribution deviates from the power-law model, inducing a performance degradation in the K sparse approximation approach as the number of coefficients increases (or the approximation error decreases). This can also be seen in Fig. 5, which shows that, though the power law fit is excellent for most cases where the fitting coefficient is above 0.99, the fit for destination port histograms of trace $\mathcal{C}$ and $\mathcal{D}$, though still very good with fitting coefficient better than 0.9, is not as good. One possible way to circumvent this limitation is to introduce more parameters in the approximation. For example, with different scaling parameters, one may model the histogram using *double power-law* distribution [13] to bound both the upper and lower tails of the reordered histogram. We let further analysis of this more sophisticated model in future work.

Finally, Fig. 8 shows that the entropy-based approach only gives a single point. In some cases, while it gives very few coefficients (small K values), the corresponding approximation error can be very high. For example, with the entropy-based approach only one coefficient in average is required to approximate the traffic per source port histogram in trace $\mathcal{A}$, which results in a σ_K close to 18%. Our approach suggests, on the other hand, that extracting 7 coefficients, for example, preserves the low dimensionality of the approximated histogram while reducing σ_K to 1%. Without an explicit relationship between K and σ_K, it is challenging to use the entropy-based approach, particularly when the approximation error is a concern. In contrast, K-sparse approximation provides such a relationship enabling to easily dimension K when a certain approximation error is targeted.

B. Impact of Sampling

To reduce the processing and storage overhead, today's commercial routers apply random packet sampling. Typically, the sampling rate can be as low as 0.1% in large, e.g, GEANT2, and medium size networks.

So far we evaluated the proposed K-sparse approximation approach using sampled traffic flow data. In this section, we investigate the validity of our approach on non-sampled flow data and the impact of sampling on the power-law observation and as a consequence on the choice of K. To this end, we use an one-month dataset of *non-sampled* traffic flow data from a lightly loaded campus network at the Norwegian University of Science and Technology (NTNU). We applied random traffic sampling at a rate of 0.1% on the collected dataset, then we investigated the behavior of the traffic histograms as well as the approximation quality of the number K of coefficients.

Fig. 9 shows the impact of sampling on source and destination port traffic histograms over a 24-hour period for the NTNU dataset. The figure illustrates that while sampling slightly shortens the tail of the distribution, the histogram after sampling matches closely with the histogram without sampling. In addition both histograms exhibit an approximately

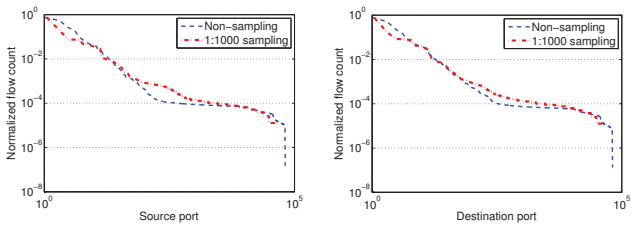
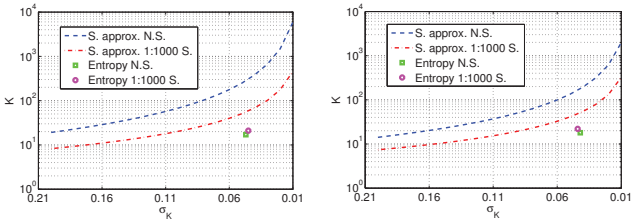


Fig. 9. Sorted traffic histograms over a 24h period (NTNU)



(a) Source port histograms (b) Destination port histograms

Fig. 10. The K value as a function of the approximation error (NTNU)

linear decay in the log-log scale, implying that power-law is a plausible fit for histograms under both sampling and non-sampling. Sampling does *not* violate the assumption of a power law decay of the reordered traffic histograms.

In Fig. 10 we plot the value K resulting from the proposed approach, when varying the approximation error within [0.21,0.01], and that from the entropy-based approach with respectively no sampling and with a sampling rate of 0.1%. Fig. 10 indicates that a histogram of sampled traffic generally requires fewer top coefficients than non-sampled traffic to achieve the same approximation error. Our investigation exposes the increase of the scaling parameter with sampling as the main reason for this observation. For example the average scaling parameter for non sampled srcport traffic histograms over the month of measurement, is in the order of 1.09, while in the order of 1.41 for sampled srcport traffic histograms. An increase in the scaling parameter induces more compressible histograms thus lower number of coefficients are required for approximation. On the other hand, the performance of the entropy-based approach is not affected by sampling, e.g, around 20 coefficient achieve 4% approximation error.

Similarly to Fig. 8, Fig. 10 finally depicts that our approach allows to control the trade-off between the allowable error and the desired compression level, while the entropy-based approach does not.

VII. Related work

Traffic histograms have been often used for traffic analysis. However they suffer from the curse of dimensionality issue. This is challenging to deal with and has thus witnessed some commercial and research activities. Several approaches investigated this issue, setting intuitively a threshold (5-10% of the total amount of traffic) to infer significant resource consumption, the top takers or applications over the network [9] [3] [4]. While this has been shown effective to reach their target, they have ignored the discussion about their choice to that particular threshold, making their analysis purely empirical. Other approaches have faced the issue for anomaly detection [12] [10] [6]. To this end, the authors of [10] have proposed for traffic histogram analysis to keep the well known source and destination ports, and remove the components that remain constant, which are associated to unused feature values, to reduce the dimensionality of the considered traffic histograms; in [10], the authors further apply Principal Component Analysis (PCA) to transform the traffic histogram into another basis where the top few components are used for the histogram approximation under that basis. While their approach was shown able to reduce the dimension of traffic histograms over source and destination ports around 10 times, it seems missing various traffic anomalies which use or target random unknown port numbers. On the other hand, the authors of [6], have proposed an aggregation strategy using hash functions to reduce traffic histogram's dimension. Their approach was shown promising providing a lossless compression technique for traffic histogram analysis. However, it suffers from a serious weakness: a mapping between the hash function and the original histogram is required, which adds an additional non negligible processing overhead. Tangentially related, the authors of [12] propose sample entropy technique to summarize traffic histograms into one value. Entropy has been widely adopted for traffic anomaly detection; however it was shown to coarsely model the properties of traffic histograms thus ineffective to detect a wide range of traffic anomalies [10]. An extension of sample entropy using relative uncertainty, has been proposed and thoroughly discussed in [15]. We believe that K sparse approximation complements these approaches while overcoming their weakness providing the missing link between the "information" and "loss" in traffic histogram dimensionality reduction.

VIII. Conclusion

In this paper, we propose K-sparse approximation: a novel and effective traffic histogram dimensionality reduction technique that uses only a small number of coefficients to approximate large traffic histograms. It is based on the observation that ordered traffic histograms of source and destination ports decay according to a power law. This has enabled us to derive an explicit relationship between K and the approximation error. We evaluate our technique using several traces from different locations and show that the power-law observation is consistent and that a very small number of coefficients are sufficient to approximate histograms with a reasonably small error. Compared to the entropy-based technique, K-sparse approximation is more flexible as it enables to directly control the trade-off between the allowable error and the desired compression level. We have also investigated our technique under sampling and found that sampling does not violate our power-law observation. These findings illustrate that K-sparse approximation is a promising technique for traffic histogram dimensionality reduction, which we believe will motivate its application and further use.

REFERENCES

[1] Geant2. http://www.geant2.net.

[2] Least square fitting. http://mathworld.wolfram.com/LeastSquaresFitting.html.

[3] Netflow analyzer. http://www.manageengine.com/products/netflow/index.html?gclid=CPan2e-a-K4CFUx0mAod6zvgzA.

[4] Solar winds. http://www.solarwinds.com/.

[5] R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[6] D. Brauckohff and K. S. X. Dimitrpoulos, A. Wagner. Anomaly extraction in backbone networks using association rules. In *ACM IMC*, 2009.

[7] A. Callado, C. Kamienski, G.Szabo, B. Gero, J. Kelner, S. Fernandes, and D. Sadok. A survey on internet traffic identification. *IEEE Communications Surveys and Tutorials*, 11(3), 2009.

[8] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM*, 51(4), 2009.

[9] C. Estan, S. Savage, and G. Varghese. Automatically inferring patterns of resource consumption in network traffic. In *ACM SIGCOMM*, 2003.

[10] A. Kind, M. Stoecklin, and X. Dimitropoulos. Histogram-based traffic anomaly detection. *IEEE Transactions on Network and Service Management*, 6(2):12–24, 2009.

[11] M. Kolosovskiy and E. Kryuchkova. Network congestion control using netflow. Technical report, Altai State Technical University, Russia, November 2009.

[12] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *ACM SIGCOMM*, 2005.

[13] A. Toda. The double power law in income distribution: Explanations and evidence. *Journal of Economic Behavior & Organization*, 2012.

[14] S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon. Providing public intradomain traffic matrices to the research community. *ACM SIGCOMM Computer Communication Review*, 36(1):83–86, 2006.

[15] K. Xu, Z. L. Zhang, and S. Bhattacharyya. Internet traffic behavior profiling for network security monitoring. *IEEE Transactions on Networking*, 16(6):1241–1252, 2008.