Sketch *-metric: Comparing Data Streams via Sketching

Emmanuelle Anceaume, Yann Busnel

Sketch ⋆-metric: Comparing Data Streams via Sketching

Emmanuelle Anceaume
IRISA / CNRS
Rennes, France
Emmanuelle.Anceaume@irisa.fr

Yann Busnel
LINA / Université de Nantes
Nantes, France
Yann.Busnel@univ-nantes.fr

***Abstract*—In this paper, we consider the problem of estimating the distance between any two large data streams in small-space constraint. This problem is of utmost importance in data intensive monitoring applications where input streams are generated rapidly. These streams need to be processed on the fly and accurately to quickly determine any deviance from nominal behavior. We present a new metric, the *Sketch ⋆-metric*, which allows to define a distance between updatable summaries (or sketches) of large data streams. An important feature of the *Sketch ⋆-metric* is that, given a measure on the entire initial data streams, the *Sketch ⋆-metric* preserves the axioms of the latter measure on the sketch. Extensive experiments conducted on both synthetic traces and real data sets allow us to validate the robustness and accuracy of the *Sketch ⋆-metric*.**

I. Introduction

The main objective of this paper is to propose a novel metric that reflects the relationships between any two discrete probability distributions in the context of massive data streams. Specifically, this metric, designated as *Sketch ⋆-metric* in the following, allows us to efficiently estimate a broad class of distances measures between any two large data streams by computing these distances only using compact synopses or sketches of the streams. The *Sketch ⋆-metric* is distribution-free and makes no assumption about the underlying data volume. It is thus capable of comparing any two data streams, identifying their correlation if any, and more generally, it allows us to acquire a deep understanding of the structure of the input streams. Formalization of this metric is the first contribution of this paper.

The interest of estimating distances between any two data streams is important in data intensive applications. Many different domains are concerned by such analyses including machine learning, data mining, databases, information retrieval, and network monitoring. In all these applications, it is necessary to quickly and precisely process a huge amount of data. For instance, in IP network management, the analysis of input streams will allow us to rapidly detect the presence of anomalies or intrusions when changes in the communication patterns occur [20]. Actually, the problem of detecting changes or outliers in a data stream is similar to the problem of identifying patterns that do not conform to the expected behavior, which has been an active area of research for many decades. To accurately analyze streams of data, a panel of information-theoretic measures and distances have been proposed to answer the specificities of the analyses. Among them, the most commonly used are the Kullback-Leibler (KL) divergence [19], or more generically, the f-divergences, introduced by Csiszar, Morimoto and Ali & Silvey [1], [15], [22], the Jensen-Shannon divergence and the Battacharyya distance [8]. Unfortunately, computing information theoretic measures of distances in the data stream model is challenging essentially because one needs to process a huge amount of data sequentially, on the fly, and by using very little storage with respect to the size of the stream. In addition the analysis must be robust over time to detect any sudden change in the observed streams (which might be the manifestation of routers deny of service attack or worm propagation). We tackle this issue by presenting an approximation algorithm that constructs a sketch of the stream from which the *Sketch ⋆-metric* is computed. This algorithm is a one-pass algorithm. It uses very basic computations, little storage space (*i.e.*, $\mathcal{O}(t(\log n + k \log m))$ where k and t are precision parameters, and m and n are respectively the size of the input stream and the number of items in the stream), and does not need any information on the structure of the input stream. This constitutes the second contribution of the paper.

Finally, the robustness of our approach is validated with a detailed experimentation study based on both synthetic traces that range from stable streams to highly skewed ones, and real data sets.

The paper is organized as follows. First, Section II reviews the related work on classical generalized metrics. Section III describes the data stream model. Section IV presents the necessary background that makes the paper self-contained. Section V formalizes the *Sketch ⋆-metric*. Section VI presents the algorithm that approximates the *Sketch ⋆-metric* in one pass and Section VII presents extensive experiments of our algorithm. Finally, we conclude in Section VIII.

II. Related Work

Work on data stream analysis mainly focuses on efficient methods (data-structures and algorithms) to answer different kind of queries over massive data streams. Mostly, these methods consist in deriving statistic estimators over the data stream, in creating summary representations of streams (to build histograms, wavelets, and quantiles), and in comparing

data streams. Regarding the construction of estimators, a seminal work is due to Alon *et al.* [2]. The authors have proposed estimators of the frequency moments F_k of a stream, which are important statistical tools that allow to quantify specificities of a data stream. Subsequently, a lot of attention has been paid to the strongly related notion of the entropy of a stream, and all notions based on entropy [14]. These notions are essentially related to the quantification of the amount of randomness of a stream (*e.g*, [5], [10], [18], [21]). The construction of synopses or sketches of the data stream have been proposed for different applications (*e.g*, [11], [12]). Actually in [17], the authors propose a characterization of the information divergences that are not sketchable. They have proven that any distance that has not "norm-like" properties is not sketchable.

Distance and divergence measures are key measures in statistical inference and data processing problems [7]. There exists two largely used broad classes of measures, namely the f-divergences and the Bregman divergences that are very important to quantify the amount of information that separates two distributions. Our goal in this paper is to formalize a metric that allows to efficiently and accurately estimate a broad class of distances measures between any two large data streams by computing these distances only on compact synopses or sketches of streams.

III. Data Stream Model

We consider a system in which a node P receives a very large data stream $\sigma = a_1, a_2, \ldots, a_m$ of data items that arrive sequentially. In the following, we describe a single instance of P, but clearly multiple instances of P may co-exist in a system (*e.g.*, in case P represents a router, or a base station in a sensor network). Each data item a_i of the stream σ is drawn from the universe $\Omega = \{1, 2, \ldots, n\}$ where n should be very large. Data items can be repeated multiple times in the stream. In the following, we suppose that the length m of the stream is not known. Items in the stream arrive regularly and quickly, and due to memory constraints, need to be processed sequentially and in an online manner. Therefore, node P can locally store only a small fraction of the items and perform simple operations on them. The algorithms we consider in this work are characterized by the fact that they can approximate some function on σ with a very limited amount of memory. We refer the reader to [23] for a detailed description of data streaming models and algorithms.

IV. Information Divergence of Data Streams

We first present notations and background that make this paper self-contained.

A. Preliminaries

- A natural approach to study a data stream σ is to model it as an empirical data distribution over the universe Ω, given by $(p_1, p_2, \ldots, p_n)$ with $p_i = x_i/m$, and $x_i = |\{j : a_j = i\}|$ representing the number of times data item i appears in σ. Note that x_i, the number of times item i appears in a stream, is commonly called the frequency of i. We have $m = \sum_{i \in \Omega} x_i$.
- 2-universal Hash Functions A collection $\mathcal{H}$ of hash functions $h : \{1, \ldots, M\} \to \{0, \ldots, M'\}$ is said to be *2-universal* if for every $h \in \mathcal{H}$ and for every two different items $i, j \in [M]$, $\mathbb{P}\{h(i) = h(j)\} \leq \frac{1}{M'}$, which is exactly the probability of collision obtained if the hash function assigned truly random values to any $i \in [M]$, where notation $[M]$ means $\{1, \ldots, M\}$.

B. Metrics and divergences

1) Metric definitions: The classical definition of a metric is based on a set of four axioms.

Definition 1 (Metric) *Given a set X, a metric is a function $d\colon X \times X \to \mathbb{R}$ such that, for any $x, y, z \in X$, we have:*

$$\textit{Non-negativity:} \quad d(x, y) \geq 0 \tag{1}$$

$$\textit{Identity of indiscernibles:} \quad d(x, y) = 0 \Leftrightarrow x = y \tag{2}$$

$$\textit{Symmetry:} \quad d(x, y) = d(y, x) \tag{3}$$

$$\textit{Triangle inequality:} \quad d(x, y) \leq d(x, z) + d(z, y) \tag{4}$$

In the context of information divergence, usual distance functions are not precisely metric. Indeed, most of divergence functions do not verify the 4 axioms, but only a subset of them. For instance, a pseudometric is a function that verifies the axioms of a metric with the exception of the identity of indiscernible, while a premetric is a pseudometric that relax both the symmetry and the triangle inequality axioms.

Two classes of generalized metrics, usually denoted as *divergences*, that allow to measure the separation of distributions have been proposed, namely the class of f-divergences and the class of Bregman divergences. Note that in the following by abusing the notation, we denote "$|\Omega|$-point distribution" by "Ω-point distribution".

2) f-divergence: The class of f-divergences provides a set of relations that is used to measure the "distance" between two distributions p and q. Mostly used in the context of statistics and probability theory, a f-divergence $\mathcal{D}_f$ is a premetric that guarantees monotonicity and convexity.

Definition 2 (f-divergence) *Let p and q be two Ω-point distributions. Given a convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$, the f-divergence of q from p is*

$$\mathcal{D}_f(p||q) = \sum_{i \in \Omega} q_i f\left(\frac{p_i}{q_i}\right),$$

where by convention $0f(\frac{0}{0}) = 0$, $af(\frac{0}{a}) = a \lim_{u \to 0} f(u)$, and $0f(\frac{a}{0}) = a \lim_{u \to \infty} f(u)/u$ if these limits exist.

Property 3 (Monotonicity) *Given κ an arbitrary transition probability that respectively transforms two Ω-point distributions p and q into p_κ and q_κ, we have:*

$$\mathcal{D}_f(p||q) \geq \mathcal{D}_f(p_\kappa||q_\kappa).$$

Property 4 (Convexity) *Let p_1, p_2, q_1 and q_2 be four Ω-point distributions. Given any $\lambda \in [0,1]$, we have:*

$$\begin{aligned}\mathcal{D}_f\left(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2\right)\\ \leq \lambda\mathcal{D}_f(p_1||q_1) + (1-\lambda)\mathcal{D}_f(p_2||q_2).\end{aligned}$$

3) Bregman divergence: Initially proposed in [9], the Bregman divergences are a generalization of the notion of distance between points. This class of generalized metrics always satisfies the non-negativity and identity of indecernibles. However they do not always satisfy the triangle inequality and their symmetry depends on the choice of the differentiable convex function F. Specifically,

Definition 5 (Bregman divergence (BD)) *Given a continuously-differentiable and strictly convex function F defined on a closed convex set C, the Bregman divergence of p from q is*

$$\mathcal{B}_F(p||q) = F(p) - F(q) - \langle \nabla F(q), (p-q)\rangle\,.$$

where the operator $\langle\cdot,\cdot\rangle$ denotes the inner product, and $\nabla F(q)$ is the gradient of F at q.

In the context of data stream, it is possible to reformulate this definition as follows. Specifically,

Definition 6 (Decomposable BD)
Let p and q be any two Ω-point distributions. Given a strictly convex function $F : (0,1] \to \mathbb{R}$, the Bregman divergence of q from p is defined as

$$\mathcal{B}_F(p||q) = \sum_{i\in\Omega}\left(F(p_i) - F(q_i) - (p_i - q_i)F'(q_i)\right).$$

The Bregman divergence verifies non-negativity and convexity properties in its first argument, but not necessarily in the second argument. Another interesting property is given by thinking of the Bregman divergence as an operator of the function F.

Property 7 (Linearity) *Let F_1 and F_2 be any two strictly convex and differentiable functions. Given any $\lambda \in [0,1]$, we have that*

$$\mathcal{B}_{F_1+\lambda F_2}(p||q) = \mathcal{B}_{F_1}(p||q) + \lambda\mathcal{B}_{F_2}(p||q).$$

4) Classical metrics: Based on these definitions, we present several commonly used metrics in Ω-point distribution context. These specific metrics are used in the evaluation part presented in Section VII.

Kullback-Leibler divergence: The Kullback-Leibler (KL) divergence [19], also called the relative entropy, is a robust metric for measuring the statistical difference between two data streams. The KL divergence owns the special feature that it is both a f-divergence and a Bregman one (with $f(t) = F(t) = t\log t$).

Given p and q two Ω-point distributions, the Kullback-Leibler divergence is defined as

$$\mathcal{D}_{KL}(p||q) = \sum_{i\in\Omega} p_i \log\frac{p_i}{q_i}. \tag{5}$$

Jensen-Shannon divergence: The Jensen-Shannon divergence (JS) is a symmetrized version of the Kullback-Leibler divergence. Also known as information radius (IRad) or total divergence to the average, it is defined as

$$\mathcal{D}_{JS}(p||q) = \frac{1}{2}D_{KL}(p||\ell) + \frac{1}{2}D_{KL}(q||\ell), \tag{6}$$

where $\ell = \frac{1}{2}(p+q)$. Note that the square root of this divergence is a metric.

Bhattacharyya distance: The Bhattacharyya distance is derived from his proposed measure of similarity between two multinomial distributions, also known as the Bhattacharya coefficient (BC) [8]. It is a semimetric as it does not verify the triangle inequality. It is defined as

$$\mathcal{D}_B(p||q) = -\log(BC(p,q)) \text{ where } BC(p,q) = \sum_{i\in\Omega}\sqrt{p_i q_i}.$$

V. Sketch $\star$-metric

We now present a method to sketch two input data streams σ_1 and σ_2, and to compute any generalized metric ϕ between these sketches such that this computation preserves all the properties of ϕ computed on σ_1 and σ_2.

Definition 8 (Sketch $\star$-metric) *Let p and q be any two Ω-point distributions. Given a precision parameter k, and any generalized metric ϕ on the set of all Ω-point distributions, there exists a Sketch $\star$-metric $\widehat{\phi}_k$ defined as follows*

$$\widehat{\phi}_k(p||q) = \max_{\rho\in\mathcal{P}_k(\Omega)} \phi(\widehat{p}_\rho||\widehat{q}_\rho),$$

with $\forall a \in \rho$, $\widehat{p}_\rho(a) = \sum_{i\in a} p_i$ and where $\mathcal{P}_k(\Omega)$ is the set of all partitions of Ω into exactly k nonempty and mutually exclusive cells.

Remark 9 *Note that for $k > n$, it does not exist a partition of Ω into k nonempty parts. By convention, we consider that $\widehat{\phi}_k(p||q) = \phi(p||q)$ in this specific context.*

In this section, we focus on the preservation of axioms and properties of a generalized metric ϕ by the corresponding Sketch $\star$-metric $\widehat{\phi}_k$.

A. Axioms preserving

Theorem 10 *Given any generalized metric ϕ then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ preserves all the axioms of ϕ.*

Proof: The proof is directly derived from Lemmata 16–19 in the companion paper [6]. The first three ones say that using sets operations and sum we get that *(i)* from non-negative numbers it is impossible to generate negative numbers, *(ii)* 0 always remains 0, and *(iii)* it is impossible to generate asymmetry. Finally, the triangle inequality is preserved as there exists $\overline{\rho} \in \mathcal{P}_k(\Omega)$ a k-cell partition such that $\phi(\widehat{p}_{\overline{\rho}}||\widehat{q}_{\overline{\rho}}) = \max_{\rho \in \mathcal{P}_k(\Omega)} \phi(\widehat{p}_\rho||\widehat{q}_\rho)$. Due to space constraints, the interested readers are invited to look through [6] for detailed proofs of these four technical lemmata. ■

B. Properties preserving

Theorem 11 *Given a f-divergence ϕ then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ is also a f-divergence.*

Proof: From Theorem 10, $\widehat{\phi}_k$ preserves the axioms of the generalized metric. Thus, $\widehat{\phi}_k$ and ϕ are in the same equivalence class. Moreover, from Lemma 13, $\widehat{\phi}_k$ verifies the monotonicity property. Thus, as the f-divergence is the only class of decomposable information *monotonic* divergences (*cf.* [15]), $\widehat{\phi}_k$ is also a f-divergence. ■

Theorem 12 *Given a Bregman divergence ϕ then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ is also a Bregman divergence.*

Proof: From Theorem 10, $\widehat{\phi}_k$ preserves the axioms of the generalized metric. Thus, $\widehat{\phi}_k$ and ϕ are in the same equivalence class. Moreover, the Bregman divergence is characterized by the property of transitivity (*cf.* [16]) defined as follows. Given p, q and r three Ω-point distributions such that $q = \Pi(L|r)$ and $p \in L$, with Π is a selection rule according to the definition of Csiszár in [16] and L is a subset of the Ω-point distributions, we have the Generalized Pythagorean Theorem:

$$\phi(p||q) + \phi(q||r) = \phi(p||r).$$

Moreover the authors in [4] show that the set S_n of all discrete probability distributions over n elements $(\{x_1, \ldots, x_n\})$ is a Riemannian manifold, and it owns another different dually flat affine structure. They also show that these dual structures give rise to the generalized Pythagorean theorem. This is verified for the coordinates in S_n and for the dual coordinates [4]. Combining these results with the projection theorem [16], [4], we obtain that

$$\begin{aligned}
\widehat{\phi}_k(p||r) &= \max_{\rho \in \mathcal{P}_k(n)} \phi(\widehat{p}_\rho||\widehat{r}_\rho) \\
&= \max_{\rho \in \mathcal{P}_k(n)} \left(\phi(\widehat{p}_\rho||\widehat{q}_\rho) + \phi(\widehat{q}_\rho||\widehat{r}_\rho)\right) \\
&= \max_{\rho \in \mathcal{P}_k(n)} \phi(\widehat{p}_\rho||\widehat{q}_\rho) + \max_{\rho \in \mathcal{P}_k(n)} \phi(\widehat{q}_\rho||\widehat{r}_\rho) \\
&= \widehat{\phi}_k(p||q) + \widehat{\phi}_k(q||r)
\end{aligned}$$

Finally, by the characterization of Bregman divergence through transitivity [16], and reinforced with Lemma 15 statement, $\widehat{\phi}_k$ is also a Bregman divergence. ■

In the following, we show that the Sketch $\star$-metric preserves the properties of divergences.

Lemma 13 (Monotonicity) *Given any generalized metric ϕ verifying the Monotonicity property then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ preserves the Monotonicity property.*

Proof: Let p and q be any two Ω-point distributions. Given $c < n$, consider a partition $\mu \in \mathcal{P}_c(\Omega)$. As ϕ is monotonic, we have $\phi(p||q) \geq \phi(\widehat{p}_\mu||\widehat{q}_\mu)$ [3]. We split the proof into two cases:

Case (1). Suppose that $c \geq k$. Computing $\widehat{\phi}_k(\widehat{p}_\mu||\widehat{q}_\mu)$ amounts in considering only the k-cell partitions $\rho \in \mathcal{P}_k(\Omega)$ that verify

$$\forall b \in \mu, \exists a \in \rho : b \subseteq a.$$

These partitions form a subset of $\mathcal{P}_k(\Omega)$. The maximal value of $\phi(\widehat{p}_\rho||\widehat{q}_\rho)$ over this subset cannot be greater than the maximal value over the whole $\mathcal{P}_k(\Omega)$. Thus we have

$$\widehat{\phi}_k(p||q) = \max_{\rho \in \mathcal{P}_k(\Omega)} \phi(\widehat{p}_\rho||\widehat{q}_\rho) \geq \widehat{\phi}_k(\widehat{p}_\mu||\widehat{q}_\mu).$$

Case (2). Suppose now that $c < k$. By definition, we have $\widehat{\phi}_k(\widehat{p}_\mu||\widehat{q}_\mu) = \phi(\widehat{p}_\mu||\widehat{q}_\mu)$. Consider $\rho' \in \mathcal{P}_k(\Omega)$ such that $\forall a \in \rho', \exists b \in \mu, a \subseteq b$. It then exists a transition probability that respectively transforms $\widehat{p}_{\rho'}$ and $\widehat{q}_{\rho'}$ into $\widehat{p}_\mu$ and $\widehat{q}_\mu$. As ϕ is monotonic, we have

$$\begin{aligned}
\widehat{\phi}_k(p||q) &= \max_{\rho \in \mathcal{P}_k(\Omega)} \phi(\widehat{p}_\rho||\widehat{q}_\rho) \\
&\geq \phi(\widehat{p}_{\rho'}||\widehat{q}_{\rho'}) \\
&\geq \phi(\widehat{p}_\mu||\widehat{q}_\mu) = \widehat{\phi}_k(\widehat{p}_\mu||\widehat{q}_\mu).
\end{aligned}$$

Finally for any value of c, $\widehat{\phi}_k$ guarantees the monotonicity property. This concludes the proof. ■

Lemma 14 (Convexity) *Given any generalized metric ϕ verifying the Convexity property then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ preserves the Convexity property.*

Proof: Let p_1, p_2, q_1 and q_2 be any four Ω-point distributions. Given any $\lambda \in [0, 1]$, we have:

$$\begin{aligned}
&\widehat{\phi}_k\left(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2\right) \\
&= \max_{\rho \in \mathcal{P}_k(\Omega)} \phi\left(\lambda \widehat{p1}_\rho + (1-\lambda)\widehat{p2}_\rho||\lambda \widehat{q1}_\rho + (1-\lambda)\widehat{q2}_\rho\right)
\end{aligned}$$

Let $\overline{\rho} \in \mathcal{P}_k(\Omega)$ such that

$$\begin{aligned}
&\phi\left(\lambda \widehat{p1}_{\overline{\rho}} + (1-\lambda)\widehat{p2}_{\overline{\rho}}||\lambda \widehat{q1}_{\overline{\rho}} + (1-\lambda)\widehat{q2}_{\overline{\rho}}\right) \\
&= \max_{\rho \in \mathcal{P}_k(\Omega)} \phi\left(\lambda \widehat{p1}_\rho + (1-\lambda)\widehat{p2}_\rho||\lambda \widehat{q1}_\rho + (1-\lambda)\widehat{q2}_\rho\right).
\end{aligned}$$

Input: Two input streams σ_1 and σ_2; the distance ϕ, k and t settings;

Output: The distance $\hat{\phi}$ between σ_1 and σ_2

Choose t functions $h : [n] \to [k]$, each from a 2-universal hash function family;
$C_{\sigma_1}[1...t][1...k] \leftarrow 0$, $C_{\sigma_2}[1...t][1...k] \leftarrow 0$;
for $a_j \in \sigma_1$ **do**
  $v \leftarrow a_j$;
  for $i = 1$ **to** t **do**
    $C_{\sigma_1}[i][h_i(v)] \leftarrow C_{\sigma_1}[i][h_i(v)] + 1$;
for $a_j \in \sigma_2$ **do**
  $w \leftarrow a_j$;
  for $i = 1$ **to** t **do**
    $C_{\sigma_2}[i][h_i(w)] \leftarrow C_{\sigma_2}[i][h_i(w)] + 1$;
On query $\hat{\phi}_k(\sigma_1||\sigma_2)$ **return** $\hat{\phi} = \max_{1\leq i\leq t}\phi(\mathrm{C}_{\sigma_1}[\mathrm{i}][-], \mathrm{C}_{\sigma_2}[\mathrm{i}][-])$;

Figure 1. *Sketch ⋆-metric* algorithm

As ϕ verifies the Convexity property, we have:

$$\begin{aligned}
&\widehat{\phi}_k\left(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2\right)\\
&= \phi\left(\lambda\widehat{p_1}_{\overline{\rho}} + (1-\lambda)\widehat{p_2}_{\overline{\rho}}||\lambda\widehat{q_1}_{\overline{\rho}} + (1-\lambda)\widehat{q_2}_{\overline{\rho}}\right)\\
&\leq \lambda\phi(\widehat{p_1}_{\overline{\rho}}||\widehat{q_1}_{\overline{\rho}}) + (1-\lambda)\phi(\widehat{p_2}_{\overline{\rho}}||\widehat{q_2}_{\overline{\rho}})\\
&\leq \lambda\left(\max_{\rho\in\mathcal{P}_k(\Omega)}\phi(\widehat{p_1}_{\rho}||\widehat{q_1}_{\rho})\right) + (1-\lambda)\left(\max_{\rho\in\mathcal{P}_k(\Omega)}\phi(\widehat{p_2}_{\rho}||\widehat{q_2}_{\rho})\right)\\
&= \lambda\widehat{\phi}_k(p_1||q_1) + (1-\lambda)\widehat{\phi}_k(p_2||q_2)
\end{aligned}$$

that concludes the proof. ■

Lemma 15 (Linearity) *The Sketch ⋆-metric definition preserves the Linearity property.*

Proof: For space limitation reasons, proof is presented in the companion paper [6]. ■

To summarize, we have shown that the *Sketch ⋆-metric* preserves all the axioms of a metric as well as the properties of f-divergences and Bregman divergences. We now show how to efficiently implement such a metric.

VI. Approximation algorithm

In this section, we propose an algorithm that computes the *Sketch ⋆-metric* in one pass on the stream. By definition of the metric (*cf.* Definition 8), we need to generate all the possible k-cell partitions. The number of these partitions follows the Stirling numbers of the second kind, which is equal to $S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$, where n is the size of the items universe. Therefore, $S(n,k)$ grows exponentially with n. As the generating function of $S(n,k)$ is equivalent to x^n, it is unreasonable in term of space complexity. We show in the following that generating $t = \lceil\log(1/\delta)\rceil$ random k-cell partitions, where δ is the probability of error of our randomized algorithm, is sufficient to guarantee good overall performance of our metric.

Our algorithm is inspired from the Count-Min Sketch algorithm proposed by Cormode and Muthukrishnan [13]. Specifically, the Count-Min algorithm is an (ε,δ)-approximation algorithm that solves the *frequency-estimation* problem. For any items in the input stream σ, the algorithm outputs an estimation $\hat{f}_v$ of the frequency of item v such that $\mathbb{P}\{|\hat{f}_v - f_v| > \varepsilon(m - f_v)\} < \delta$, where m represent the size of the input stream and $\varepsilon, \delta > 0$ are given as parameters of the algorithm. The estimation is computed by maintaining a two-dimensional array C of $t \times k$ counters, and by using t 2-universal hash functions h_i ($1 \leq i \leq t$), where $k = 2/\varepsilon$ and $t = \lceil\log(1/\delta)\rceil$. Each time an item v is read from the input stream, this causes one counter of each line to be incremented, *i.e.*, $C[i][h_i(v)]$ is incremented by one for each $i \in [1..t]$.

To compute the *Sketch ⋆-metric* of two streams σ_1 and σ_2, two sketches $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$ of these streams are constructed according to the above description. Note that there is no particular assumption on the length of both streams σ_1 and σ_2. That is their respective length is finite but unknown. By construction of the 2-universal hash functions h_i ($1 \leq i \leq t$), the i^{th} line of C_{σ_1} and C_{σ_2} corresponds to the same partition ρ_i of the Ω-point empirical distributions of both σ_1 and σ_2. Thus when a query is issued to compute the given distance ϕ between these two streams, the maximal value over all the t partitions ρ_i of the distance ϕ between $\widehat{\sigma}_{1_{\rho_i}}$ and $\widehat{\sigma}_{2_{\rho_i}}$ is returned, *i.e.*, the distance ϕ applied to the i^{th} lines of C_{σ_1} and C_{σ_2} for $1 \leq i \leq t$. Figure 1 presents the pseudo-code of our algorithm.

Lemma 16 *Given parameters k and t, Algorithm 1 gives an approximation of the Sketch ⋆-metric, using*

$$\mathcal{O}\left(t(\log n + k\log m)\right) \text{ bits of space.}$$

Proof: The matrices C_{σ_i}, for any $i \in \{1,2\}$, are composed of $t \times k$ counters, which uses $\mathcal{O}(\log m)$. On the other hand, with a suitable choice of hash family, we can store the hash functions above in $\mathcal{O}(t\log n)$ space. ■

VII. Performance Evaluation

A. Settings of the experiments

We have implemented our *Sketch ⋆-metric* and have conducted a series of experiments on different types of streams and for different parameters settings. We have fed our algorithm with both real-world data sets and synthetic traces. Real data give a realistic representation of some existing systems, while the latter ones allow to capture phenomenon which may be difficult to obtain from real-world traces, and thus allow to check the robustness of our metric. We have varied all the significant parameters of our algorithm, that is, the maximal number of distinct data items n in each stream, the number of cells k of each generated partition,

Table I
STATISTICS OF THE FIVE REAL DATA TRACES.

Data trace	# items (m)	# distinct items (n)	max. freq.
NASA (July)	1,891,715	81,983	17,572
NASA (August)	1,569,898	75,058	6,530
ClarkNet (August)	1,654,929	90,516	6,075
ClarkNet (September)	1,673,794	94,787	7,239
Saskatchewan	2,408,625	162,523	52,695

and the number of generated partitions t. Synthetic traces of streams have been generated from 7 distributions showing very different shapes, that is the Uniform distribution, the Zipfian or power law one with parameter $\alpha = 1, 2, 4$, the Poisson distribution with parameter $\lambda = N/2$, the Binomial and the Negative Binomial (or Pascal) ones. For each parameters setting, we have conducted and averaged 100 trials of the same experiment, leading to a total of more than $300,000$ experiments for the evaluation of our metric. Real data have been downloaded from the repository of Internet network traffic [24]. We have used five large traces among the available ones. Two of them represent two weeks logs of HTTP requests to the Internet service provider ClarkNet WWW server – ClarkNet is a full Internet access provider for the Metro Baltimore-Washington DC area – the other two ones contain two months of HTTP requests to the NASA Kennedy Space Center WWW server, and the last one represents seven months of HTTP requests to the WWW server of the University of Saskatchewan, Canada. Table I presents some statistics of these data traces. Note that all these benchmarks share a Zipfian behavior, with a lower α for the University of Saskatchwan.

B. Main lessons drawn from the experiments

In this section, we evaluate the accuracy of the *Sketch $\star$-metric* by comparing $\widehat{\phi}_k(p||q)$ with $\phi_k(p||q)$, for $\phi \in \{$Kullback-Leiber, Jensen-Shannon, Bhattacharyya$\}$, and for p and q generated from the 7 distributions and the 5 real data sets. Distances computed from the sketches of the stream are referred to as *Sketch* in the legend of the graphs, while the ones computed from the full streams are mentioned as *Ref.* Due to space constraints, only a subset of the results are presented in the paper. The interested reader is invited to read [6] for the complete evaluation.

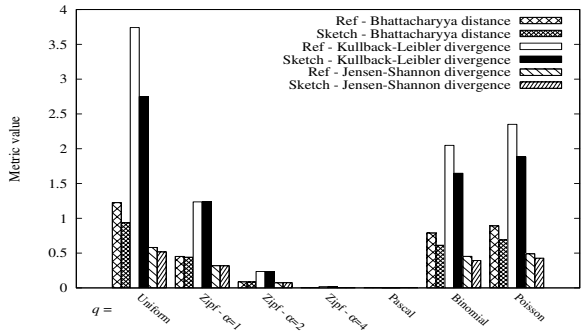
Figure 2 shows the accuracy of our metric as a function of the different input streams and the different generalized metrics applied on these streams. The first noticeable remark is that *Sketch $\star$-metric* behaves perfectly well when the two compared streams follow the same distribution, whatever the generalized metric ϕ used. This can be observed from both synthetic traces (*cf.* Figure 2(a) with both p and q following the Pascal distribution, and Figure 2(b) with both p and q uniformly distributed), and real data sets (*cf.* Figures 2(c) and 2(d) with the NASA (July and August) and ClarkNet (August and September) traces).

This tendency is further observed when the distributions of input streams are close to each other (*e.g.*, Zipf–$\alpha = 2, 4$ and Pascal distributions, or Uniform and Zipf–$\alpha = 1$). This makes the *Sketch $\star$-metric* a very good candidate as a parametric method for making distribution parameters inference. Another interesting result is shown when the two input streams exhibit a totally different shape. Specifically, let us consider Figures 2(a) and 2(b). Sketching the Uniform distribution leads to k-cell partitions whose value is well distributed, that is, for a given partition ϕ, all the k cell values have with high probability the same value. Now, when sketching the Pascal distribution, the repartition of the data items in the cells of any given partitions is such that a few number of data items (those with high frequency) populate a very few number of cells. However, the values of these cells is very large compared to the other cells, which are populated by a large number of data items whose frequency is small. Thus, the contribution of data items exhibiting a small frequency and sharing the cells of highly frequent items is biased compared to the contribution of the other items. Thus although the input streams show a totally different shape, the accuracy of $\widehat{\phi}_k$ is only slightly lowered in these scenarios which makes it a very powerful tool to compare any two different data streams. The same observation holds with real data sets. When the shapes of the input streams are different (which is the case for Saskatchewan with respect to the 4 other input streams), the accuracy of the *Sketch $\star$-metric* decreases a little bit but in a very small proportion. Notice that the scales on the y-axis differ significantly in Figures 2(a)-2(b) and in Figures 2(c)-2(d).

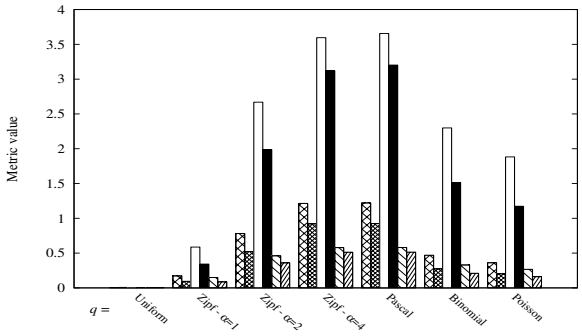
We have also observed in [6] the strong impact of the non-symmetry of the Kullback-Leibler divergence on the computation of the distance (computed on full streams or on sketches) with a clear influence when the input streams follow a Pascal and Zipf–$\alpha = 1$ distributions.

Figure 3 summarizes the good properties of $\widehat{\phi}_k$ by illustrating how, for any generalized metric ϕ, and for any variations in the shape of the two input distributions, $\widehat{\phi}_k$ remains close to ϕ. Recall that increasing values of the r parameter of the Negative Binomial distribution makes the shape of the distribution flatter, while maintaining the same mean value.

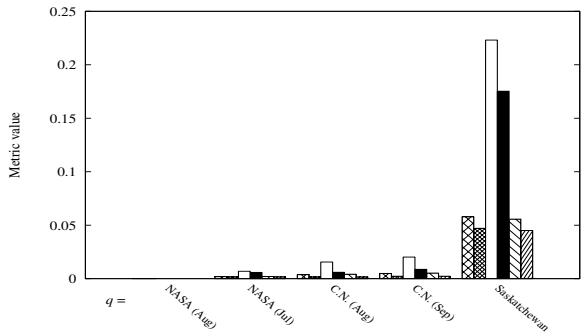
Figure 4 presents the impact of the number of cells per generated partition on the accuracy of the $\star$-metric on both synthetic traces and real data. It clearly shows that by increasing k the number of data items per cell in the generated partition shrinks and thus the absolute error on the computation of the distance decreases. The same feature appears when the number n of distinct data items in the stream increases. Indeed, when n increases (for a given k), the number data items per cell augments and thus the precision of our metric decreases. This gives rise to a shift of the inflection point, as illustrated in Figure 4(b) as data sets have almost twenty to forty times more distinct data items than the synthetic ones. As aforementioned, the



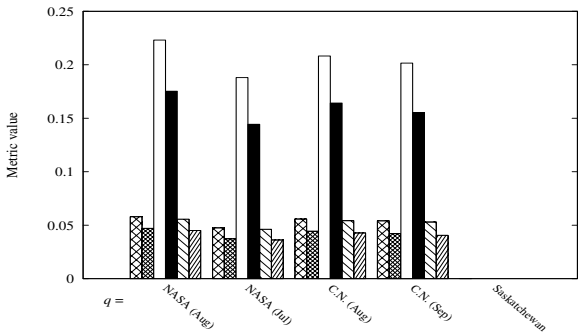
(a) Synthetic traces – Distribution p follows a Negative Binomial NB(3; 0.99) (or Pascal) distribution



(b) Synthetic traces – Distribution p follows a Uniform distribution



(c) Real datasets – The input stream p is the NASA (August) trace



(d) Real datasets – The input stream p is the Saskatchewan trace

Figure 2. Comparison between the *Sketch ⋆-metric* and the ϕ metric as a function of the input stream q either generated from a distribution or real traces. For synthetic traces, $m = 200,000$ and $n = 4,000$. Parameters of the count-min sketch data structure are $k = 200$ and $t = 4$. All the histograms share the same legend, but for readability reasons, this legend is only indicated on histogram 2(a).

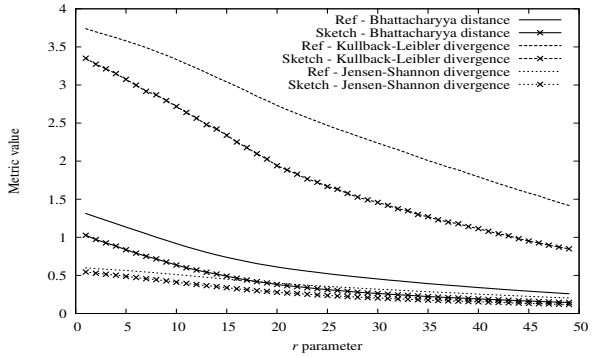


Figure 3. Comparison between the *Sketch ⋆-metric* and the ϕ metric as a function of the parameters of the Negative Binomial distribution NB$(r, n/(2r+n))$, where distribution p follows a Uniform distribution and q follows the Negative Binomial distribution NB$(r, n/(2r+n))$.

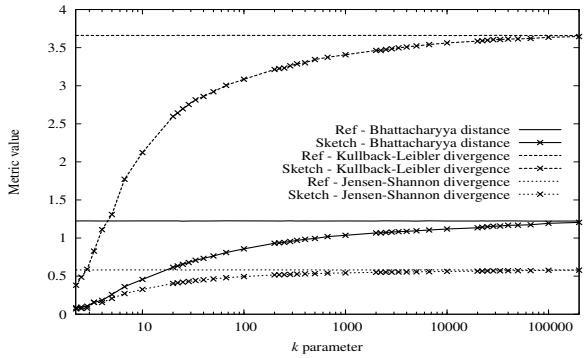
input streams exhibit very different shapes which explain the strong impact of k. Note also that k has the same influence on the *Sketch ⋆-metric* for all the generalized distances ϕ.

Finally, it is interesting to note that the number t of generated partitions has a slight influence on the accuracy of our metric The reason comes from the use of 2-universal hash functions, which guarantee for each of them and with high probability that data items are uniformly distributed over the cells of any partition. As a consequence, augmenting the number of such hash functions has a weak influence on the accuracy of the metric.

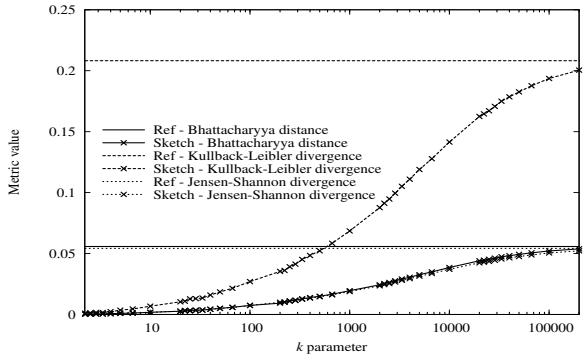
VIII. Conclusion and open issues

In this paper, we have introduced a new metric, the *Sketch ⋆-metric*, that allows to compute any generalized metric ϕ on the summaries of two large input streams. We have presented a simple and efficient algorithm to sketch streams and compute this metric, and we have shown that it behaves pretty well whatever the considered input streams. We are convinced of the undisputable interest of such a metric in various domains including machine learning, data mining, databases, information retrieval and network monitoring.

Regarding future works, we plan to consider a distributed setting, where each site would be in charge of analyzing its own streams and then would propagate its results to the other sites of the system for comparison or merging. An immediate application of such a tool would be to detect massive attacks in a decentralized manner (*e.g.*, by identifying specific connection profiles as with worms propagation, and massive port scan attacks or by detecting sudden variations in the volume of received data).



(a) Synthetic traces – Distribution p follows a Uniform distribution and q follows a Negative Binomial NB$(3; 0.99)$ one



(b) Real datasets – The input stream p is the ClarkNet (August) trace and q is the Saskatchewan one

Figure 4. Comparison between the *Sketch $\star$-metric* and the ϕ metric as a function of the number of cells k per partition (the number of partitions t of the count-min sketch data structure is set to 4). For synthetic traces, $m = 200,000$ and $n = 4,000$.

REFERENCES

[1] S. M. Ali and S. D. Silvey. General Class of Coefficients of Divergence of One Distribution from Another. *J. of the Royal Statistical Society. Series B.*, 28(1):131–142, 1966.

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proc. of the 28th ACM STOC*, pages 20–29, 1996.

[3] S.-I. Amari. α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, nov 2009.

[4] S.-I. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

[5] E. Anceaume and Y. Busnel. An information divergence estimation over data streams. In *Proc. of the 11th IEEE Intl Symp. on Network Computing and Applications (NCA)*, 2012.

[6] E. Anceaume and Y. Busnel. Sketch $\star$-metric: Comparing Data Streams via Sketching. Technical Report hal-00721211, CNRS / LINA, 2012.

[7] M. Basseville and J.-F. Cardoso. On entropies, divergences, and mean values. In *Proc. of the IEEE International Symposium on Information Theory*, 1995.

[8] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bul. of the Calcutta Math. Soc.*, 35:99–109, 1943.

[9] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[10] A. Chakrabarti, K. D. Ba, and S. Muthukrishnan. Estimating entropy and entropy norm on data streams. In *Proc. of the 23rd Intl STACS*. Springer, 2006.

[11] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. *TCS*, 312(1):3–15, 2004.

[12] G. Cormode and M. Garofalakis. Sketching probabilistic data streams. In *Proc. of the 2007 ACM SIGMOD intl conf. on Management of Data*, pages 281–292, 2007.

[13] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.

[14] T. Cover and J. Thomas. Elements of information theory. *Wiley New York*, 1991.

[15] I. Csiszár. Information Measures: A Critical Survey. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 73–86, Dordrecht, 1978. D. Riedel.

[16] I. Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

[17] S. Guha, P. Indyk, and A. Mcgregor. Sketching information divergences. *Machine Learning*, 72(1-2):5–19, 2008.

[18] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proc. of the 17th Annual ACM-SIAM SODA*, pages 733–742, 2006.

[19] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[20] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *Proc. of the ACM Conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, 2005.

[21] A. Lall, V. Sekar, M. Ogihara, J. Xu, and H. Zhang. Data streaming algorithms for estimating entropy of network traffic. In *Proc. of the SIGMETRICS*. ACM, 2006.

[22] T. Morimoto. Markov processes and the h-theorem. *J. of the Physical Society of Japan*, 18(3):328–331, 1963.

[23] Muthukrishnan. *Data Streams: Algorithms and Applications*. Now Publishers Inc., 2005.

[24] the Internet Traffic Archive. http://ita.ee.lbl.gov/html/traces.html. Lawrence Berkeley National Laboratory, Apr. 2008.