HAL open science

Sketch -metric: Comparing Data Streams via Sketching

Emmanuelle Anceaume, Yann Busnel

▶ To cite this version:

Emmanuelle Anceaume, Yann Busnel. Sketch -metric: Comparing Data Streams via Sketching. 2013, pp.8. hal-00764772

HAL Id: hal-00764772
https://hal.science/hal-00764772

Submitted on 13 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Publications Internes de l'IRISA
ISSN : 2102-6327
PI 2001 – Décembre 2012

Sketch ⋆-metric: Comparing Data Streams via Sketching

Emmanuelle Anceaume* Yann Busnel**

Abstract: 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 (such as the non-negativity, the identity, the symmetry, the triangle inequality but also specific properties of the f-divergence or the Bregman one). Extensive experiments conducted on both synthetic traces and real data sets allow us to validate the robustness and accuracy of the *Sketch ⋆-metric*.

Key-words: Data stream; metric; randomized approximation algorithm.

Sketch ⋆-metrique: Comparaison de flots de donnes base sur des rsums ("sketch")

Résumé : *Nous étudions le problème li l'estimation de la distance entre de flots de données quelconques sous hypothèse de calcul et mmoire limitée. Ce problme s'avère être très important dans les applications de monitoring où les flots de données sont générés rapidement.*

Mots clés : *Flots de données, algorithme d'approximation randomizé.*

---

* CNRS UMR 6074 IRISA, emmanuelle.anceaume@irisa.fr, CIDRE
** LINA, Université de Nantes, Yann.Busnel@univ-nantes.fr, ATLAS-GDD

©IRISA – Campus de Beaulieu – 35042 Rennes Cedex – France – +33 2 99 84 71 00 – www.irisa.fr

1 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* $\star$*-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* $\star$*-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 [8]. 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 [27]. In sensors networks, such an analysis will enable us to determine any correlation between geographical and environmental informations [12]. 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. For instance, depending on the specificities of the domain considered and the type of outliers considered, different methods have been designed, namely classification-based, clustering-based, nearest neighbor based, statistical, spectral, and information theory. 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 [26], or more generically, the f-divergences, introduced by Csiszar, Morimoto and Ali & Silvey [19, 29, 1], the Jensen-Shannon divergence and the Battacharyya distance [10]. More details can be found in the comprehensive survey of Basseville [9]. 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 may 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* $\star$*-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 2 reviews the related work on classical generalized metrics and their applications on the data stream model while Section 3 describes this model. Section 4 presents the necessary background that makes the paper self-contained. Section 5 formalizes the *Sketch* $\star$*-metric*. Section 6 presents the algorithm that fairly approximates the *Sketch* $\star$*-metric* in one pass and Section 7 presents extensive experiments (on both synthetic traces and real data sets) of our algorithm. Finally, we conclude in Section 8.

2 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 (*i.e.*, norm and relative entropy) [18]. These notions are essentially related to the quantification of the amount of randomness of a stream (*e.g*, [14, 23, 13, 28, 5, 24, 33]). The construction of synopses or sketches of the data stream have been proposed for different applications (*e.g*, [15, 17, 16]).

Distance and divergence measures are key measures in statistical inference and data processing problems [9]. There exists two largely used broad classes of measures, namely the f-divergences and the Bregman divergences. Among them, there exists two classical distances, namely the Kullback-Leibler (KL) divergence and the Hellinger distance, that are very important to quantify the amount of information that separates two distributions. In [7], the authors have proposed a one pass algorithm for estimating the KL divergence of an observed stream compared to an expected one. Experimental evaluations have shown that the estimation provided by this algorithm is accurate for different adversarial settings for which the quality of other methods dramatically decreases. However, this solution assumes that the expected stream is the uniform one, that is a fully random stream. Actually in [22], 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.

Our goal in this paper is to go one step further by formalizing 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.

3 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 [30] for a detailed description of data streaming models and algorithms.

4 Information Divergence of Data Streams

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

4.1 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. We have $m = \sum_{i \in \Omega} x_i$.

4.1.1 Entropy

Intuitively, the entropy is a measure of the randomness of a data stream σ. The entropy $H(\sigma)$ is minimum (*i.e.*, equal to zero) when all the items in the stream are the same, and it reaches its maximum (*i.e.*, $\log_2 m$) when all the items in the stream are distinct. Specifically, we have $H(\sigma) = -\sum_{i \in \Omega} p_i \log_2 p_i$. In the following, the log is to the base 2 and thus entropy is expressed in bits. By convention, we have $0 \log 0 = 0$. Note that the number of times x_i item i appears in a stream is commonly called the frequency of i.

4.1.2 2-universal Hash Functions

In the following, we intensively use hash functions randomly picked from a 2-universal hash family. 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 $\{1, \ldots, M\}$.

4.2 Metrics and divergences

4.2.1 Metric definitions

The classical definition of a metric is based on a set of four axioms.

Definition 4.1 (Metric) *Given a set X, a **metric** is a function $d : X \times X \to \mathbb{R}$ 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. We recall hereafter some definitions of generalized metrics.

Definition 4.2 (Pseudometric) *Given a set X, a **pseudometric** is a function that verifies the axioms of a metric with the exception of the* identity of indiscernible*, which is replaced by*

$$\forall x \in X, d(x, x) = 0.$$

Note that this definition allows that $d(x, y) = 0$ for some $x \neq y$ in X.

Definition 4.3 (Quasimetric) *Given a set X, a **quasimetric** is a function that verifies all the axioms of a metric with the exception of the* symmetry *(*cf. *Relation 3).*

Definition 4.4 (Semimetric) *Given a set X, a **semimetric** is a function that verifies all the axioms of a metric with the exception of the* triangle inequality *(*cf. *Relation 4).*

Definition 4.5 (Premetric) *Given a set X, a **premetric** is a pseudometric that relax both the* symmetry *and* triangle inequality *axioms.*

Definition 4.6 (Pseudoquasimetric) *Given a set X, a **pseudoquasimetric** is a function that relax both the* identity of indiscernible *and the* symmetry *axioms.*

Note that the latter definition simply corresponds to a premetric satisfying the triangle inequality. Remark also that all the generalized metrics preserve the *non-negativity* axiom.

4.2.2 Divergences

We now give the definition of two broad classes of generalized metrics, usually denoted as *divergences*.

f-divergence 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 4.7 (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.

Following this definition, any f-divergence verifies both monotonicity and convexity.

Property 4.8 (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 4.9 (Convexity) *Let p_1, p_2, q_1 and q_2 be four Ω-point distributions. Given any $\lambda \in [0, 1]$, we have:*

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

This class of divergences has been introduced in independent works by Csiszár, Morimoto and Ali & Silvey [19, 29, 1], in chronological order. All the distance measures in the so-called *Ali-Silvey distances* are applicable to quantifying statistical differences between data streams.

Bregman divergence Initially proposed in [11], this class of generalized metrics encloses quasimetrics and semimetrics, as these divergences do not satisfy the triangle inequality nor symmetry.

Definition 4.10 (Bregman divergence (BD)) *Given F a continuously-differentiable and strictly convex function defined on a closed convex set C, the **Bregman divergence** associated with F for $p, q \in C$ is defined as*

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

where the operator $\langle \cdot, \cdot \rangle$ denotes the inner product.

In the context of data stream, it is possible to reformulate this definition according to probability theory. Specifically,

Definition 4.11 (Decomposable BD)
*Let p and q be two Ω-point distributions. Given a strictly convex function $F : (0, 1] \to \mathbb{R}$, the **Bregman divergence** associated with F of q from p is defined as*

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

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

Property 4.12 (Linearity) *Let F_1 and F_2 be two strictly convex and differentiable functions. Given any $\lambda \in [0, 1]$, we have that*

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

4.2.3 Classical metrics

In this section, we present several commonly used metrics in Ω-point distribution context. These specific metrics are used in the evaluation part presented in Section 7.

Kullback-Leibler divergence The Kullback-Leibler (KL) divergence [26], 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 then defined as

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

where $H(p) = -\sum p_i \log p_i$ is the (empirical) entropy of p and $H(p,q) = -\sum p_i \log q_i$ is the cross entropy of p and q.

Jensen-Shannon divergence The Jensen-Shannon divergence (JS) is a symmetrized and smoothed 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) [10]. It is defined as

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

This distance is a semimetric as it does not verify the triangle inequality. Note that the famous Hellinger distance [25] equal to $\sqrt{1 - BC(p,q)}$ verifies it.

5 Sketch $\star$-metric

Given this context, 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. Proof of correctness of this method is presented in this section.

Definition 5.1 (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*

$$\widehat{\phi}_k(p||q) = \max_{\rho \in \mathcal{P}_k(\Omega)} \phi(\widehat{p}_\rho || \widehat{q}_\rho) \text{ with } \forall a \in \rho,\ \widehat{p}_\rho(a) = \sum_{i \in a} p_i,$$

where $\mathcal{P}_k(\Omega)$ is the set of all partitions of Ω into exactly k nonempty and mutually exclusive cells.

Remark 5.2 *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* $\widehat{\phi}_k$.

5.1 Axioms preserving

Theorem 5.3 *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 1 *The proof derives directly from Lemmata 5.4, 5.5, 5.6 and 5.7. The three first ones say that using sets operations and sum then,* (i) *from non-negative number it is impossible to generate negative numbers,* (ii) 0 *always remains* 0*, and* (iii) *it is impossible to generate asymmetry.*

Lemma 5.4 (Non-negativity) *Given any generalized metric ϕ verifying the Non-negativity axiom then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ preserves the Non-negativity axiom.*

Proof 2 *Let p and q be any two Ω-point distributions. By definition,*

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

As for any two k-point distributions, ϕ is positive we have $\widehat{\phi}_k(p||q) \geq 0$ that concludes the proof.

Lemma 5.5 (Identity of indiscernible) *Given any generalized metric ϕ verifying the Identity of indiscernible axiom then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ preserves the Identity of indiscernible axiom.*

Proof 3 *Let p be any Ω-point distribution. We have*

$$\widehat{\phi}_k(p||p) = \max_{\rho \in \mathcal{P}_k(\Omega)} \phi(\widehat{p}_\rho||\widehat{p}_\rho) = 0,$$

due to ϕ Identity of indiscernible axiom.

Consider now two Ω-point distributions p and q such that $\widehat{\phi}_k(p||q) = 0$. Metric ϕ verifies both the non-negativity axiom (by construction) and the Identity of indiscernible axiom (by assumption). Thus we have $\forall \rho \in \mathcal{P}_k(\Omega), \widehat{p}_\rho = \widehat{q}_\rho$, leading to

$$\forall \rho \in \mathcal{P}_k(\Omega), \forall a \in \rho, \sum_{i \in a} p(i) = \sum_{i \in a} q(i). \tag{7}$$

Moreover, for any $i \in \Omega$, there exists a partition $\rho \in \mathcal{P}_k(\Omega)$ such that $\{i\} \in \rho$. By Equation 7, $\forall i \in \Omega, p(i) = q(i)$, and so $p = q$.

Combining the two parts of the proof leads to $\widehat{\phi}_k(p||q) = 0 \Longleftrightarrow p = q$, which concludes the proof of the Lemma.

Lemma 5.6 (Symmetry) *Given any generalized metric ϕ verifying the Symmetry axiom then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ preserves the Symmetry axiom.*

Proof 4 *Let p and q be any two Ω-point distributions. We have*

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

Let $\overline{\rho} \in \mathcal{P}_k(\Omega)$ be a k-cell partition such that $\phi(\widehat{p}_{\overline{\rho}}||\widehat{q}_{\overline{\rho}}) = \max_{\rho \in \mathcal{P}_k(\Omega)} \phi(\widehat{p}_\rho||\widehat{q}_\rho)$. We get

$$\widehat{\phi}_k(p||q) = \phi(\widehat{p}_{\overline{\rho}}||\widehat{q}_{\overline{\rho}}) = \phi(\widehat{q}_{\overline{\rho}}||\widehat{p}_{\overline{\rho}}) \leq \widehat{\phi}_k(q||p).$$

By symmetry, considering $\underline{\rho} \in \mathcal{P}_k(\Omega)$ such that $\phi(\widehat{q}_{\underline{\rho}}||\widehat{p}_{\underline{\rho}}) = \max_{\rho \in \mathcal{P}_k(\Omega)} \phi(\widehat{q}_\rho||\widehat{p}_\rho)$, we also have $\widehat{\phi}_k(q||p) \leq \widehat{\phi}_k(p||q)$, which concludes the proof.

Lemma 5.7 (Triangle inequality) *Given any generalized metric ϕ verifying the Triangle inequality axiom then, for any $k \in \mathbb{N}$, the corresponding Sketch $\star$-metric $\widehat{\phi}_k$ preserves the Triangle inequality axiom.*

Proof 5 *Let p, q and r be any three Ω-point distributions. Let $\overline{\rho} \in \mathcal{P}_k(\Omega)$ be a k-cell partition such that $\phi(\widehat{p}_{\overline{\rho}}||\widehat{q}_{\overline{\rho}}) = \max_{\rho \in \mathcal{P}_k(\Omega)} \phi(\widehat{p}_\rho||\widehat{q}_\rho)$. We have*

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

that concludes the proof.

5.2 Properties preserving

Theorem 5.8 *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 6 *From Theorem 5.3, $\widehat{\phi}_k$ preserves the axioms of the generalized metric. Thus, $\widehat{\phi}_k$ and ϕ are in the same equivalence class. Moreover, from Lemma 5.10, $\widehat{\phi}_k$ verifies the monotonicity property. Thus, as the f-divergence is the only class of decomposable information* monotonic *divergences (*cf. *[20]), $\widehat{\phi}_k$ is also a f-divergence.*

Theorem 5.9 *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 7 *From Theorem 5.3, $\widehat{\phi}_k$ 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. *[21]) 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 [21] 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 [21, 4], we obtain that

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

Finally, by the characterization of Bregman divergence through transitivity [21], and reinforced with Lemma 5.12 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 5.10 (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 8 *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 5.11 (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 9 *Let p_1, p_2, q_1 and q_2 be any four Ω-point distributions. Given any $\lambda \in [0, 1]$, we have:*

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

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

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

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 5.12 (Linearity) *The Sketch $\star$-metric definition preserves the Linearity property.*

Algorithm 1: Sketch $\star$-metric algorithm

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] → [k], each from a 2-universal hash function family;
C_σ1[1...t][1...k] ← 0;
C_σ2[1...t][1...k] ← 0;
for a_j ∈ σ1 do
    v ← a_j;
    for i = 1 to t do
        C_σ1[i][h_i(v)] ← C_σ1[i][h_i(v)] + 1;
for a_j ∈ σ2 do
    w ← a_j;
    for i = 1 to t do
        C_σ2[i][h_i(w)] ← C_σ2[i][h_i(w)] + 1;
On query φ̂_k(σ1||σ2) return φ̂ = max_{1≤i≤t} φ(C_σ1[i][−], C_σ2[i][−]);
```

Proof 10 *Let F_1 and F_2 be two strictly convex and differentiable functions, and any $\lambda \in [0,1]$. Consider the three Bregman divergences generated respectively from F_1, F_2 and $F_1 + \lambda F_2$.*

Let p and q be two Ω-point distributions. We have:

$$\begin{aligned}\widehat{\mathcal{B}}_{F_1+\lambda F_{2k}}(p||q) &= \max_{\rho\in\mathcal{P}_k(\Omega)} \mathcal{B}_{F_1+\lambda F_2}(\widehat{p}_\rho||\widehat{q}_\rho)\\ &= \max_{\rho\in\mathcal{P}_k(n)} (\mathcal{B}_{F_1}(\widehat{p}_\rho||\widehat{q}_\rho) + \lambda\mathcal{B}_{F_2}(\widehat{p}_\rho||\widehat{q}_\rho))\\ &\leq \widehat{\mathcal{B}}_{F_{1k}}(p||q) + \lambda\widehat{\mathcal{B}}_{F_{2k}}(p||q)\end{aligned}$$

As F_1 and F_2 are two strictly convex functions, and taken a leaf out of the Jensen's inequality, we have:

$$\begin{aligned}&\widehat{\mathcal{B}}_{F_{1k}}(p||q) + \lambda\widehat{\mathcal{B}}_{F_{2k}}(p||q)\\ &\leq \max_{\rho\in\mathcal{P}_k(\Omega)} (\mathcal{B}_{F_1}(\widehat{p}_\rho||\widehat{q}_\rho) + \lambda\mathcal{B}_{F_2}(\widehat{p}_\rho||\widehat{q}_\rho))\\ &= \widehat{\mathcal{B}}_{F_1+\lambda F_{2k}}(p||q)\end{aligned}$$

that concludes the proof.

To summarize, we have shown that the *Sketch $\star$-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.

6 Approximation algorithm

In this section, we propose an algorithm that computes the *Sketch $\star$-metric* in one pass on the stream. By definition of the metric (*cf.* Definition 5.1), 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 [17]. Specifically, the Count-Min algorithm is an (ε,δ)-approximation algorithm that solves the *frequency-estimation*

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

Table 1: Statistics of real data traces.

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 $\mathbb{P}\{|\hat{f}_v - f_v| > \varepsilon f_v\} < \delta$, where $\varepsilon, \delta > 0$ 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.*, $C[h_i(v)]$ 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 $(1 \leq i \leq t)$, each line of C_{σ_1} and C_{σ_2} corresponds to one 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 ith 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 6.1 *Given parameters k and t, Algorithm 1 gives an approximation of the Sketch ⋆-metric, using*

$$\mathcal{O}\left(t(\log n + k \log m)\right)$$

bits of space.

Proof 11 *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.*

7 Performance Evaluation

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, and the number of generated partitions t. 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 [32]. 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. In the following these data sets will be respectively referred to as ClarkNet, NASA, and Saskatchewan traces. Table 1 presents some statistics of these data traces, in term of stream size (*cf.* "# items" in the table), number of distinct items in each stream (*cf.* "# distinct") and the number of occurrences of the most frequent item (*cf.* "max. freq."). Figure 1 illustrates the shape of each real data set distribution. Note that all these benchmarks share a Zipfian behavior, with a lower α parameter for the University of Saskatchwan.

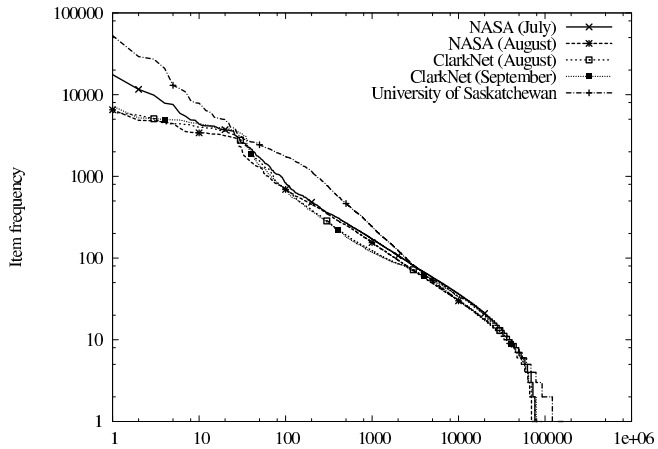


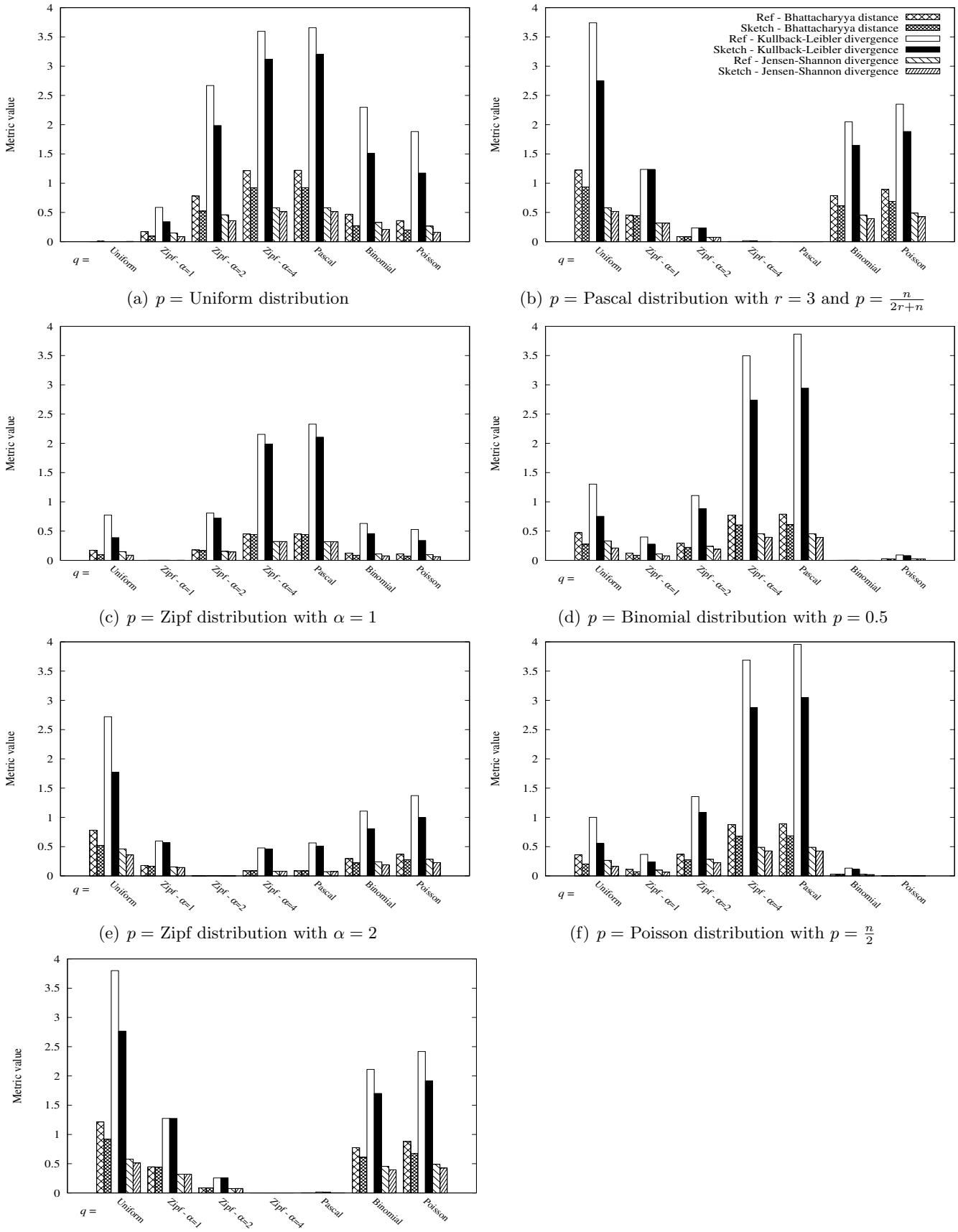
Figure 1: Log-log scale distribution of frequencies for each real data trace.

We have evaluated the accuracy of our metric by comparing for each data set (real and synthetic), the results obtained with our algorithm on the stream sketches (referred to as *Sketch* in the legend) and the ones obtained on full streams (referred to as *Ref* distance in the legend). That is, for each couple of input streams, and for each generalized metric ϕ, we have computed both the exact distance between the two streams and the one as generated by $\widehat{\phi}_k$. By distance between full streams, we mean that the metric has been applied on the (empirical) distribution of $|\Omega| = n$ points (versus k points used in the sketch $\star$-metric). We now present the main lessons drawn from these experiments. The reader is invited to look at the full experiments analysis provided in the companion paper [6].

Figure 2 and 3 show the accuracy of our metric as a function of the different input streams and the different generalized metrics applied on these streams. All the histograms shown in Figures 2(a)–3(b) share the same legend, but for readability reasons, this legend is only indicated on histogram 2(b). Three generalized metrics have been used, namely the Bhattacharyya distance, the Kullback-Leibler and the Jensen-Shannon divergences, and five distribution families denoted by p and q have been compared with these metrics.

Let us focus on synthetic traces. The first noticeable remark is that our metric behaves perfectly well when the two compared streams follow the same distribution, whatever the generalized metric ϕ used (*cf.*, Figure 2(a) with the uniform distribution, Figures 2(c), 2(e) and 2(g) with Zipfian distributions, Figure 2(b) with the Pascal distribution, Figure 2(d) with the Binomial distribution, and Figure 2(f) with the Poisson one). This tendency can be observed when the distributions of input streams are close (*e.g*, Zipfian distributions with different parameter α, or Pascal and Zipf with $\alpha = 4$), which makes the Sketch $\star$-metric a very good candidate as a parametric method for making inference about the parameters of the distribution that follow input streams. A more interesting result is shown when the two input distributions exhibit a totally different shape. Specifically, let us consider as input distributions the Uniform and the Pascal distributions (see Figure 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 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.

We can also observe 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 with $\alpha = 1$ distributions (see Figure 2(b) and 2(c)).



(a) $p =$ Uniform distribution

(b) $p =$ Pascal distribution with $r = 3$ and $p = \frac{n}{2r+n}$

(c) $p =$ Zipf distribution with $\alpha = 1$

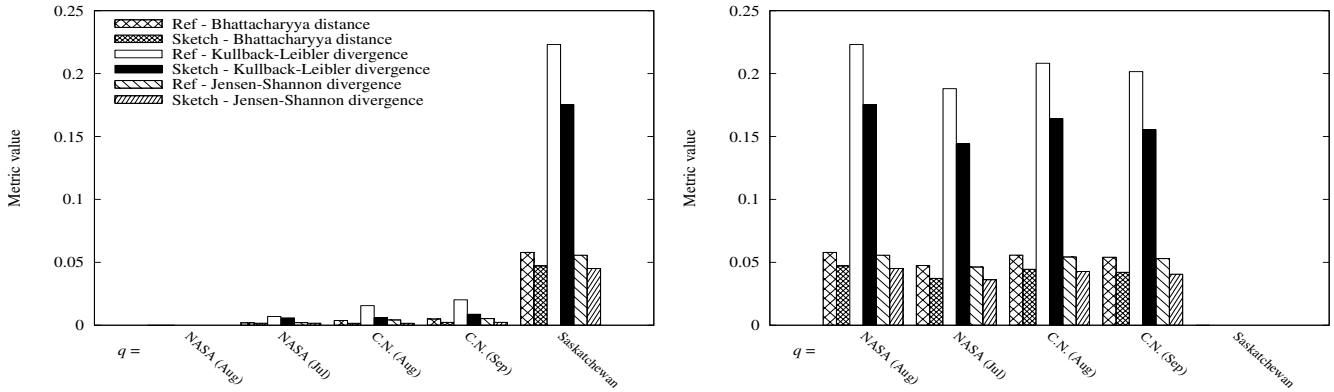
(d) $p =$ Binomial distribution with $p = 0.5$

(e) $p =$ Zipf distribution with $\alpha = 2$

(f) $p =$ Poisson distribution with $p = \frac{n}{2}$

(g) $p =$ Zipf distribution with $\alpha = 4$

Figure 2: *Sketch ⋆-metric* accuracy as a function of p and q (or r for 4). Parameters setting is as follows: $m = 200,000$; $n = 4,000$; $k = 200$; $t = 4$ where m represents the size of the stream, n the number of distinct data items in the stream, t the number of generated partitions and k the number of cells per generated partition.



(a) p = NASA webserver (August) (b) p = Saskatchewan University webserver

Figure 3: *Sketch* $\star$*-metric* accuracy as a function of real data traces. Parameters setting: $k = 2,000$; $t = 4$.

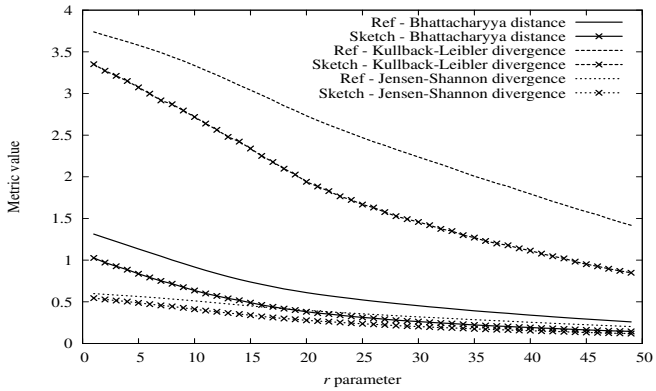
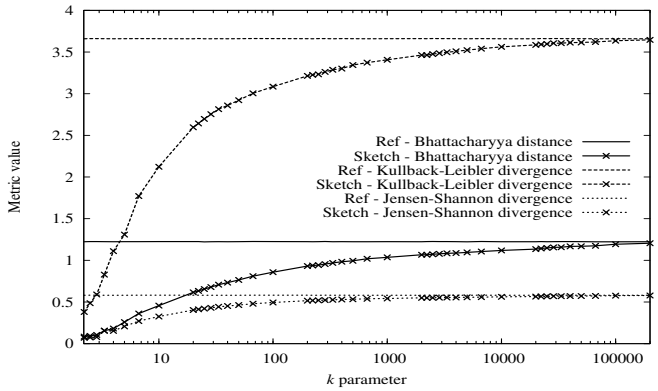


Figure 4: p = Uniform distribution and q = Pascal distribution, as a function of its parameter r $\left(\text{such that its second parameter } p = \frac{n}{2r+n}\right)$.

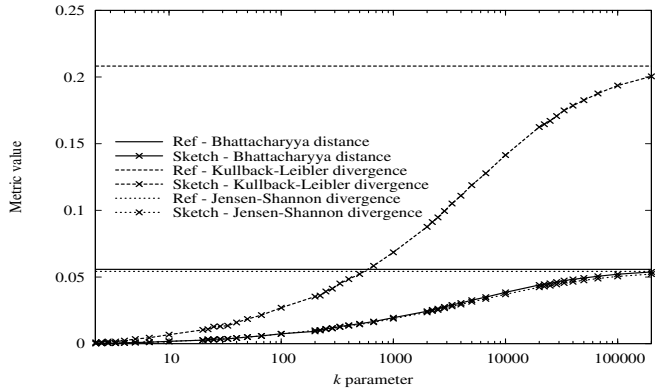
Finally, Figure 4 summarizes the good properties of our metric by illustrating how for any generalized metric ϕ, and for any variations in the shape of the two input distributions our metric $\widehat{\phi}_k$ remains close to ϕ. Recall that increasing values of the Pascal distribution parameter r – while maintaining the mean value – makes the shape of the Pascal distribution flatter.

The same general remarks hold when considering real data sets. Indeed, Figure 3 shows that when the input streams are close to each other, which is the case for both NASA (July and August) and ClarkNet (August and September) traces (*cf.* Figure 1), then applying the generalized metrics ϕ on sketches gives good results with respect to full streams. 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 Figure 2 and in Figure 3.

Figure 5 presents the impact of the number of cells per generated partition on the accuracy of our 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 5(b), due to the fact that data sets have almost twenty times more distinct data items than the synthetic ones. As aforementioned, the input streams exhibit very different shapes which explain the strong impact of k. Note also that k has the same influence on the *Sketch* $\star$*-metric* for all the generalized distances ϕ.

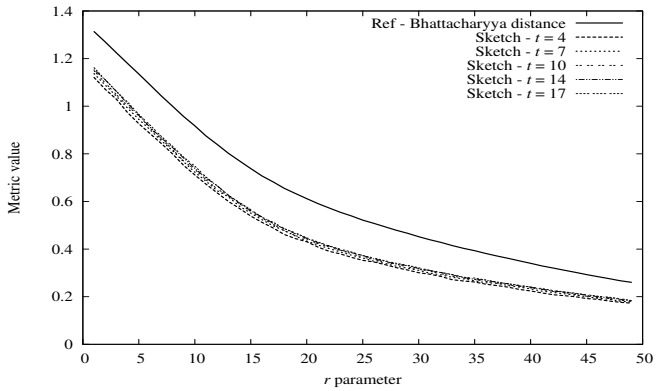


(a) *Sketch ⋆-metric* accuracy as a function of k. We have $m = 200,000$; $n = 4,000$; $t = 4$; $r = 3$

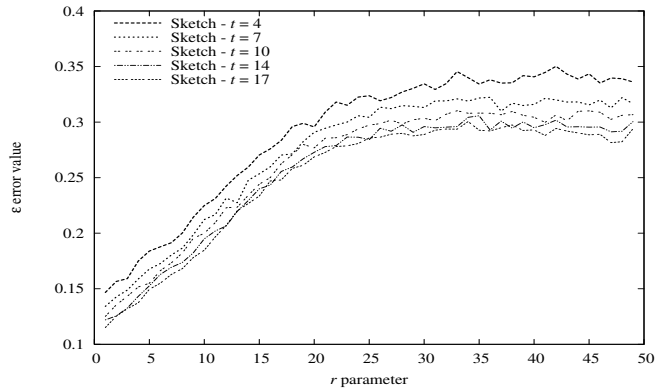


(b) Sketch ⋆-metric accuracy between data trace extracted from ClarkNetwork (August) and Saskatchewan University, as a function of k

Figure 5: *Sketch ⋆-metric* between the Uniform distribution and Pascal with parameter $p = \frac{n}{2r+n}$ (Figures 5(a)), and between data trace extracted from ClarkNetwork (August) and Saskatchewan University (Figures 5(b)).



(a) Value of Bhattacharyya distance



(b) Difference with Bhattacharyya distance

Figure 6: *Sketch ⋆-metric* estimation between Uniform distribution and Pascal with parameter $p = \frac{n}{2r+n}$, as a function of t and r.

It is interesting to note that the number t of generated partitions has a slight influence on the accuracy of our metric [6]. 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. Finally, Figure 6 presents the error made by the *Sketch ⋆-metric* for five different values of t as a function of parameter r of the Pascal distribution. Figures 6(b) depicts for each value of t the difference between the reference and the sketch values which makes more visible the impact of t. The same main lesson drawn from these figures is the moderate impact of t on the precision of our algorithm.

8 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 characterize our metric among Rényi divergences [31], also known as α-divergences, which generalize different divergence classes. We also 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).

References

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

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing (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 *Proceedings of the 11th IEEE International Symposium 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, 2012.

[7] E. Anceaume, Y. Busnel, and S. Gambs. AnKLe: detecting attacks in large scale systems via information divergence. In *Proceedings of the 9th European Dependable Computing Conference (EDCC)*, 2012.

[8] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proceedings of 21st ACM Symposium on Principles of Database Systems (PODS)*, 2002.

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

[10] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[11] 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.

[12] Y. Busnel, M. Bertier, and A.-M. Kermarrec. SOLIST or How To Look For a Needle in a Haystack? In *the 4th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob'2008)*, Avignon, France, October 2008.

[13] A. Chakrabarti, K. D. Ba, and S. Muthukrishnan. Estimating entropy and entropy norm on data streams. In *In Proceedings of the 23rd International Symposium on Theoretical Aspects of Computer Science (STACS)*. Springer, 2006.

[14] A. Chakrabarti, G. Cormode, and A. McGregor. A near-optimal algorithm for computing the entropy of a stream. In *In ACM-SIAM Symposium on Discrete Algorithms*, pages 328–335, 2007.

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

[16] G. Cormode and M. Garofalakis. Sketching probabilistic data streams. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 281–292, 2007.

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

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

[19] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutató Int. Közl*, 8:85–108, 1963.

[20] 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.

[21] 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.

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

[23] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 733–742, 2006.

[24] Z. Haung, K. Yi, and Q. Zhang. Randomized algorithms for tracking distributed count, frequencies and ranks. In *Proceedings of 31st ACM Symposium on Principles of Database Systems (PODS)*, 2012.

[25] E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *J. Reine Angew. Math.*, 136:210–271, 1909.

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

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

[28] A. Lall, V. Sekar, M. Ogihara, J. Xu, and H. Zhang. Data streaming algorithms for estimating entropy of network traffic. In *Proceedings of the joint international conference on Measurement and modeling of computer systems (SIGMETRICS)*. ACM, 2006.

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

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

[31] A. Renyi. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1960.

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

[33] B. R. Z. Liu and M. Vojnovic. Continuous distributed counting for non-monotonic streams. In *Proceedings of 31st ACM Symposium on Principles of Database Systems (PODS)*, 2012.