# PhD Proposal: Functional monitoring problem for distributed large-scale data streams

Emmanuelle Anceaume, Yann Busnel, Bruno Sericola
IRISA / CNRS Rennes
LINA / Université de Nantes
INRIA Rennes – Bretagne Atlantique

January 24, 2013

**Abstract**

In this PhD proposal, we consider the setting of large scale distributed systems, in which each node needs to quickly process a huge amount of data received in the form of a stream that may have been tampered with by an adversary. In this situation, several fundamental problems has been raised recently, that concern many domains 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. We propose to combine sampling techniques and information-theoretic methods to extract pertinent information from such a streams (metrics, summaries, pattern matching, *etc.*). Unfortunately, computing information theoretic measures 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).

On the other hand, very few works have tackled the distributed streaming model, also called the functional monitoring problem [12], which combines features of both the streaming model and communication complexity models. As in the streaming model, the input data is read on the fly, and processed with a minimum workspace and time. In the communication complexity model, each node receives an input data stream, performs some local computation, and communicates only with a coordinator who wishes to continuously compute or estimate a given function of the union of all the input streams. The challenging issue in this model is for the coordinator to compute the given function by minimizing the number of communicated bits [12, 6, 15].

**Keywords :** *Large-scale Data Stream; Randomized approximation algorithm; Functional monitoring problem; Byzantine Adversary; Performance Analysis*

## Introduction

### Context and issues

The interest of estimating metrics or identify specific patterns between several data streams is important in data intensive applications. Many different domains are concerned by such anal-

yses 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 allows to rapidly detect the presence of anomalies or intrusions when changes in the communication patterns occur.

The problem of extracting pertinent information 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. A comprehensive survey of these techniques, their advantages and their drawbacks is given in [10]. A common feature of these techniques is their space complexity and their computational cost, as they rely on full space algorithms for analyzing their data.

As a specific example, the main objective of [2] is the online estimation of the similarity between observed data streams and expected (*i.e.* idealized) ones in order to detect in real time the presence of intrusions in network traffic. More precisely, we have proposed a distributed algorithm that approximates with guaranteed error bounds in a single pass and with both a small amount of storage memory and processing capacity, the relative entropy between massive and high frequency distributed sequences of data. This works perfectly fits the IP network traffic context, however it could be applied to any other data issued from distributed applications such as social networks or sensor readings.

Given our settings — the real time monitoring of network traffic with little capacities in terms of storage and processing — relying on full space algorithms for analyzing input data is not feasible. In contrast, two main approaches exist to monitor in real time massive data streams. The first one consists in regularly sampling the input streams so that only a limited amount of data items is locally kept [23, 19, 20]. This allows to exactly compute functions on these samples. However, accuracy of this computation, with respect to the stream in its entirety, fully depends on the volume of data that has been sampled and their locations in the stream. Worse, an adversary may easily take advantage of the sampling policy to hide its attacks among packets that are not sampled, or in a way that prevents its "malicious" packets to be correlated. In contrast, the streaming approach consists in scanning each piece of data of the input stream on the fly, and in locally keeping only compact synopses or sketches that contain the most important information about data items. This approach enables to derive some data streams statistic with guaranteed error bounds without making any assumptions on the order in which data items are received at nodes (*i.e.*, data items ordering can be manipulated by an omnipotent adversary [4]). Most of the research done so far with this approach has focused on computing functions or statistic measures with error $\varepsilon$ using poly$(1/\varepsilon, \log n)$ space where $n$ is the domain size of the data items. These include the computation of the number of different data items in a given stream [7, 14, 18], the frequency moments [1], the most frequent data items [1, 11], the entropy of the stream [9, 21], or the relative entropy between one data stream and the uniform one [2, 5].

## Problems and opportunities

Unfortunately, computing information theoretic measures 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).

On the other hand, very few works have tackled the distributed streaming model, also called the functional monitoring problem [12], which combines features of both the streaming model and communication complexity models. As in the streaming model, the input data is read on the fly, and processed with a minimum workspace and time. In the communication complexity model, each node receives an input data stream, performs some local computation, and communicates only with a coordinator who wishes to continuously compute or estimate a given function of the union of all the input streams. The challenging issue in this model is for the coordinator to compute the given function by minimizing the number of communicated bits [12, 6, 15]. Cormode *et al.* [12] pioneer the formal study of functions in this model by focusing on the estimation of the first three frequency moments $F_0$, $F_1$ and $F_2$ [1]. Arackaparambil *et al.* [6] consider the empirical entropy estimation [1] and improve the work of Cormode by providing lower bounds on the frequency moments, and finally distributed algorithms for counting at any time $t$ the number of items that have been received by a set of nodes from the inception of their streams have been proposed in [17, 22].

For instance, following this model, we go a step further by proposing an estimator called AnKLe (Attack-tolerant eNhanced Kullback-Leibler divergence Estimator) that estimates the relative entropy, or the Kullback-Leibler (KL) divergence between distributed streams. This divergence can be viewed as an extension of the Shannon entropy and is often referred to as the relative entropy [13]. Citing Chakrabarti *et al.* [8], "[...] rationale of estimating entropy-based distances is that there are intimate connections between the randomness of traffic sequences (formalized as the entropy) and the propagation of malicious events. Indeed, detecting sudden changes in a stream may be a good indicator of attacks". Note that in [16], 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.

## Requested work

### Objectives

Most of the work proposed for a single stream are clearly not adaptable to the distributed functional monitoring model [12]. The concrete objective of this PhD proposal is the design and the prototypical implementation of some efficient one-pass distributed algorithms, in the context of social-based personal cloud network, where massive data stream exchanges is the norm.

Specifically, the first objective of this thesis is to propose an enhanced metric that reflects the relationships between any set of discrete probability distributions in the context of massive data streams, in order to modelize any user behavior and to provide some relationship metric nor proximity between them. This metric should be able to efficiently estimate a broad class of distances measures between large data streams by computing these distances only using compact synopses or sketches of the streams. It should be distribution-free and should make no assumption about the underlying data volume. The second step is to propose a one-pass distributed algorithm that approximates this novel metric with a given probability $\delta$. This algorithm should use very little space and few operations (i.e., sublinear in the parameters of

the system — size of the stream, number of distinct items in the streams). Implementation and comparison with previous contribution of the literature is obviously a mandatory step to validate the relevance of the solution.

Finally, the existing literature in the context of the distributed functional monitoring model does not take into account the semantic and/or the type of data that composed all the streams. This approach provides some very generic solutions but should not be optimal in specific application context [2]. As our system architecture and applications are clearly focused for this PhD, we also plan to restrict the system model and to propose some enhanced algorithm by taking into account the inherent syntax or content of these data streams.

### Working plan

The work will obviously starts by a study of the state of the art, which shares several problematics (data stream model, distributed functional monitoring problem, social networks, *etc.*). Quite early, some experimentation should be realized to compare existent solutions on different extended data sets, in order to identify which approach are optimized in our context (for instance, sampling approach as in [1, 7] in contrast of sketching ones [3, 23]). Moreover, all experimentations should be realized on the testbed composed by a 48 Rasberry Pi cluster, own by the GDD team.

This first phase should let us raised a first algorithm design and some concrete targeted application. It will then require to incrementally refine this propositions both of design and implementation.

By the way, in addition to the experimental validation of these new solutions, all proposed algorithms would be theoretically proved in the aforementioned model. This permits to extract lower and upper bounds in term of space and time complexity, and to estimate precisely the approximation error due to the probabilistic approach of this model.

## Applicants

### Skills

The applicants must have important algorithmic skills and should know about applied mathematics, such as probability and statistics.

Mastering distributed systems would be appreciate.

## References


[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the 28th annual ACM symposium on Theory of computing (STOC)*, pages 20–29, 1996.

[2] 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.

[3] E. Anceaume and Y. Busnel. Sketch *-metric: Comparing Data Streams via Sketching. Technical report, CIDER - IRISA , CIDRE - INRIA - SUPELEC , Laboratoire d'Informatique de Nantes Atlantique - LINA, 7 2012. 12 pages, double colonnes.

[4] E. Anceaume, Y. Busnel, and S. Gambs. Uniform and Ergodic Sampling in Unstructured Peer-to-Peer Systems with Malicious Nodes. In *Proceedings of the 14th international conference on Principles of distributed systems (OPODIS)*, volume 6490, pages 64–78, 2010.

[5] 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.

[6] C. Arackaparambil, J. Brody, and A. Chakrabarti. Functional monitoring without monotonicity. In *Proceedings of the 36th ACM International Colloquium on Automata, Languages and Programming: Part 1*, 2009.

[7] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques (RANDOM)*, pages 1–10. Springer-Verlag, 2002.

[8] 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.

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

[10] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

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

[12] G. Cormode, S. Muthukrishnan, and K. Yi. Algorithms for distributed functional monitoring. In *Proceedings of the 19th annual ACM-SIAM Symposium On Discrete Algorithms (SODA)*, 2008.

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

[14] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31(2):182–209, 1985.

[15] P. B. Gibbons and S. Tirthapura. Estimating simple functions on the union of data streams. In *Proceedings of the Thirteenth Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, pages 281–291, 2001.

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

[17] 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.

[18] D. M. Kane, J. Nelson, and D. P. Woodruff. An optimal algorithm for the distinct element problem. In *Proceedings of the Symposium on Principles of Databases (PODS)*, 2010.

[19] V. Karamcheti, D. Geiger, Z. Kedem, and S. Muthuskrishnan. Detecting malicious network traffic using inverse distribution of packet contents. In *Proceedings of the workshop on Mining Network Data (MineNet) co-located with ACM SICOMM*, 2005.

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

[21] 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.

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

[23] B. K. Subhabrata, E. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. Sketch-based change detection: Methods, evaluation, and applications. In *Internet Measurement Conference*, pages 234–247, 2003.