# Thesis Proposal

# Bandwidth Efficient Distributed Monitoring Schemes

Yuval Alfassi

Supervised by Prof. Daniel Keren

University of Haifa
Faculty of Social Sciences
Department of Computer Science

December, 2018

# Abstract

Distributed monitoring is a problem that arises when trying to monitor properties of dynamic data which is spread distributively. Tracking the value of a function over dynamic data in a distributed setting is a challenging problem in numerous real-world modern applications. Several monitoring schemes were proposed as an approach to coping with this problem in order to reduce the amount of communication messages between servers.

This research will focus on developing new distributed monitoring schemes which use much less communication bandwidth than existing methods. This will be done by exploiting some unique properties of convex functions so that the dimensionality of the transmitted data will be reduced.

# 1 Introduction

Monitoring a function over a large amount of dynamically changing data in a distributed fashion is a common computer-science challenge. Whether it's monitoring features of distributed sensor networks [1], top-k monitoring [2], monitoring distributed ratio queries [3] or tracking properties of large distributed dynamic graphs [4], innovative approaches had to be developed in order to deal with the difficulties of the data being both dynamic and distributed.

The need for minimizing both the bandwidth and the processing power is described in [5]; in the *Big Data* era, where data is of very high dimensionality and changes rapidly, data transmission over a communication channel has to be devised cleverly. Transmition of high dimensional data is not only extremely time consuming but also virtually impractical; for instance, a system of air pollution sensors which distributively have to determine the air pollution level may benefit from communication reduction [6].

## 1.1 Problem Definition

Commonly, the distributed monitoring model is focused on determining whether a function over dynamic distributed data crosses a certain threshold. This is used as a component to the *distributed function approximation problem* [7], which $\varepsilon$-approximates the value of a function over time. The distributed model is described as follows:

1. There are $n$ data-servers, $s_1 ... s_n$
2. A central *coordinator*, $c$ exists, with whom the servers communicate over a communication channel.
3. $server_i$ knows only its dynamic data – the local vector, $v_i$.
4. The global data, represented by the global vector $v$, is the average of the local vectors:

$$v = \frac{1}{n} \sum_{i=1}^{n} v_i \tag{1}$$

5. A function $f$ is monitored over the global vector $v$ so it's $\varepsilon$-approximated with 100% confidence: let the estimation be the dynamic value $\mu$ (without loss of generality, assume $\mu \geq 0$), then at all times:

$$(1-\varepsilon)\mu \leq f(v) \leq (1+\varepsilon)\mu \tag{2}$$

The *threshold monitoring problem* [7] monitors whether the function's value crosses a certain threshold. The function approximation problem is commonly reduced to two simultaneous threshold monitoring problems: let $T = (1+\varepsilon)\mu$ be the upper-bound threshold's value, then, the upper-bound monitoring objective is to determine whether:

$$f(v) \leq T \tag{3}$$

Likewise, the lower-bound function monitoring is performed with the threshold $(1-\varepsilon)\mu$.

In turn, this threshold monitoring can be treated as a *geometric monitoring problem* [8], where one tries to find a *safe zone* of vectors $\{v \mid f(v) \leq T\}$, which is convex, so every convex combination of vectors inside this safe zone is also inside the safe zone, see Fig. 1. This geometric safe zone approach is the fundamental idea behind distributed monitoring techniques.

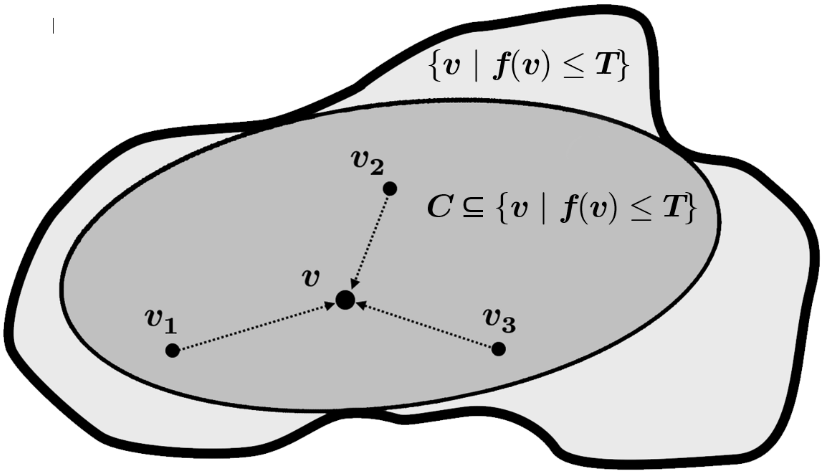

Figure 1: Convex Safe Zone

$C$ is a convex bound of the set $\{v \mid f(v) \leq T\}$; since $v_1, v_2, v_3 \in C$, so does the average vector $v \in C$.

# 2 Previous Work

## 2.1 Linear Functions

Since linear functions are additive and homogeneous, the basic algorithm for distributively monitoring their value is fairly easy. Since $f(v) = \frac{1}{n} \sum f(v_i)$, tracking the value of the global $f(v)$ isn't quite complicated. Work about linear functions such as the distributed count problem was done at [9]. However, things get more complicated when dealing with non-linear functions.

## 2.2 The Covering Spheres Method

The first approach which exploited the geometric view of the ditributed monitoring problem is the *Covering Spheres* method [8].

This method artificially creates a convex safe zone (as shown in Fig. 1) where the server's local data could be at without a need for communication.

This method seemed very effective theoratically, though it proved to be impractical. The *Covering Spheres* method demands performing lots of time consuming heavy mathematical operations, so it isn't scalable computation-wise [10].

Furthermore, the violation resolution phase demands transmitting high dimensional vectors, which makes the communication bandwidth quite too high.

## 2.3 The Convex Decomposition Method

Another distributed monitoring scheme previously developed is the *Convex Decomposition* method [11]. This method composes a convex safe zone by decomposing into half-planes the complement of the geometric space of the monitored condition. The *Convex Decomposition* method geometrically monitoring whether the average global vector is in the intersection of the half-planes.

Unfortunately, this method suffers from similar issues as the *Covering Spheres* method, and cannot be applied on some basic functions, thus isn't always feasable [10]

## 2.4 The Convex Bound Method

The *Covering Spheres* method and the *Convex Decomposition* method turned out to be impractical albeit their mathematical foundation. In need of more computationally lightweight and consistent monitoring approach, the *Convex Bound* method was proposed [10].

The *Convex Bound* method tightly bounds the monitored function by a convex function, so the convex bound serves as the convex safe zone: when monitoring $f(v) \leq T$, an upper bound convex $c$ has to be found so for all $v$, $f(v) \leq c(v)$.

And the new monitoring objective becomes:

$$c(v) \leq T \tag{4}$$

The same goes for lower bound threshold monitoring – it's done by bounding $f$ from below by a concave function.

Accordingly, the distributed function monitoring is performed on the convex bound functions, which is far simpler due to the convexity property. On the other hand, this method is also very expensive regarding the bandwidth consumption; high dimensional data is frequently transmitted on the communication channel.

# 3 Research Objectives

1. Introducing multiple innovative distributed monitoring schemes which avoid sending high dimensional data unless its crucially needed.

2. Proving the *Distance Lemma*, a lemma used as a basis of an efficient distributed monitoring scheme we'll develop. The *Distance Lemma* states that given a convex body and several points, if the sum of distances to the convex border from the points inside the convex body is greater than the sum of distances to the border from the points on the outside, then the average of the points is inside the convex body, Thus, distributed monitoring can be performed by sending just one scalar.

3. Incorporating data-sketches into distributed monitoring schemes without damaging the 0% false-negative necessity of the distributed monitoring problem.

4. Conducting several experiments, laying out comparisons of multiple attributes of distributed monitoring schemes on real-world data, focusing on the bandwidth consumption.

# References

[1] S. Burdakis and A. Deligiannakis, "Detecting outliers in sensor networks using the geometric approach," in *2012 IEEE 28th International Conference on Data Engineering.* IEEE, 2012, pp. 1108–1119.

[2] B. Babcock and C. Olston, "Distributed top-k monitoring," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data.* ACM, 2003, pp. 28–39.

[3] R. Gupta, K. Ramamritham, and M. Mohania, "Ratio threshold queries over distributed data sources," *Proceedings of the VLDB Endowment*, vol. 6, no. 8, pp. 565–576, 2013.

[4] A. McGregor, D. Tench, S. Vorotnikova, and H. T. Vu, "Densest subgraph in dynamic graph streams," in *International Symposium on Mathematical Foundations of Computer Science.* Springer, 2015, pp. 472–482.

[5] N. Giatrakos, Y. Kotidis, A. Deligiannakis, V. Vassalos, and Y. Theodoridis, "In-network approximate computation of outliers with quality guarantees," *Information Systems*, vol. 38, no. 8, pp. 1285–1308, 2013.

[6] W.-L. Cheng, Y.-C. Kuo, P.-L. Lin, K.-H. Chang, Y.-S. Chen, T.-M. Lin, and R. Huang, "Revised air quality index derived from an entropy function," *Atmospheric Environment*, vol. 38, no. 3, pp. 383–391, 2004.

[7] M. Garofalakis, D. Keren, and V. Samoladas, "Sketch-based geometric monitoring of distributed stream queries," *Proceedings of the VLDB Endowment*, vol. 6, no. 10, pp. 937–948, 2013.

[8] I. Sharfman, A. Schuster, and D. Keren, "A geometric approach to monitoring threshold functions over distributed data streams," *ACM Transactions on Database Systems (TODS)*, vol. 32, no. 4, p. 23, 2007.

[9] R. Keralapura, G. Cormode, and J. Ramamirtham, "Communication-efficient distributed monitoring of thresholded counts," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, 2006, pp. 289–300.

[10] A. Lazerson, D. Keren, and A. Schuster, "Lightweight monitoring of distributed streams," *ACM Transactions on Database Systems (TODS)*, vol. 43, no. 2, p. 9, 2018.

[11] A. Lazerson, I. Sharfman, D. Keren, A. Schuster, M. Garofalakis, and V. Samoladas, "Monitoring distributed streams using convex decompositions," *Proceedings of the VLDB Endowment*, vol. 8, no. 5, pp. 545–556, 2015.