Monitoring Distributed Streams using Convex Decompositions

Arnon Lazerson, Izchak Sharfman, Assaf Schuster. Technion I.I.T. {lazerson,tsachis,assaf}@cs.technion.ac.il
Daniel Keren. Haifa University. dkeren@cs.haifa.ac.il
Minos Garofalakis, Vasilis Samoladas. Technical University of Crete. {minos,vsam}@softnet.tuc.gr

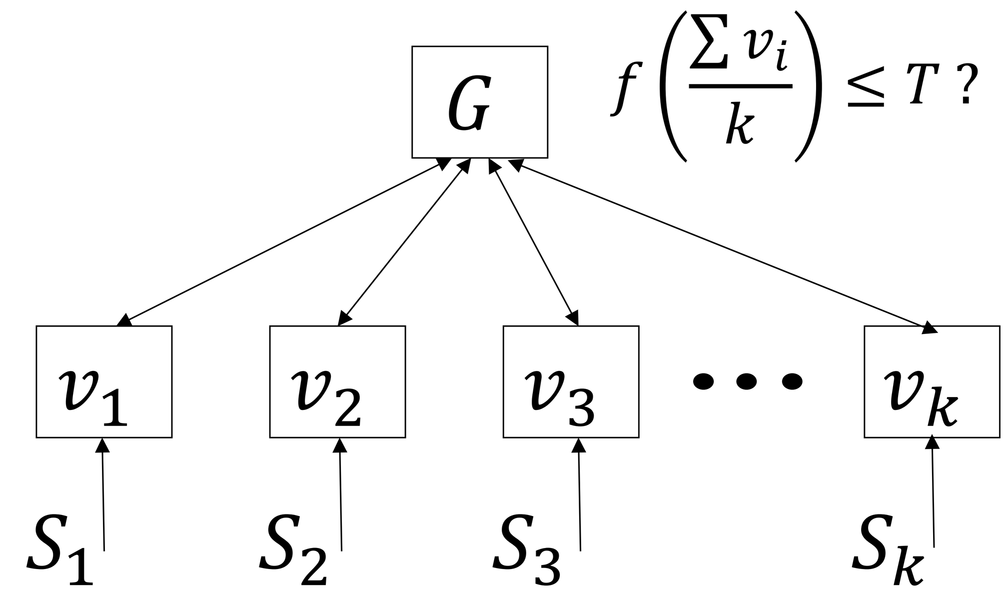
Background

Large-scale monitoring applications rely on continuous tracking of complex queries over distributed data streams. Effective distributed stream processing solutions must be

- Space efficient
- Time efficient
- **Communication efficient**

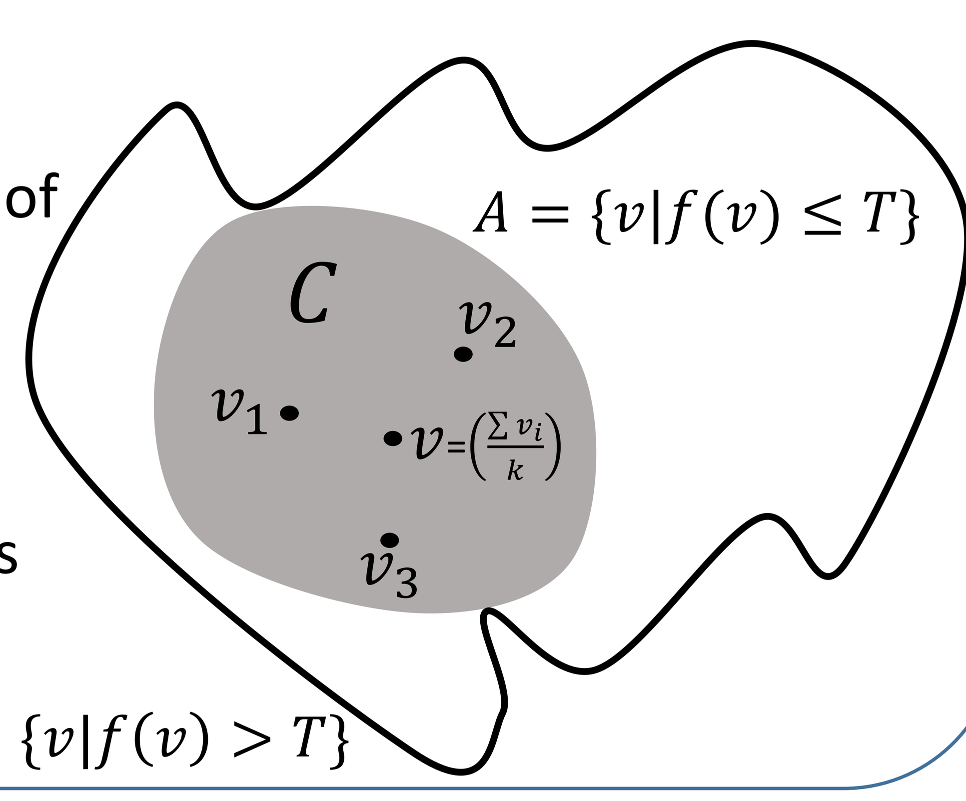
Distributed Monitoring Model

Distributed streams S_i continuously update the local vectors v_i. The coordinator G must issue an alert when the global condition $f\left(\frac{\sum v_i}{k}\right) \le T$ is breached.

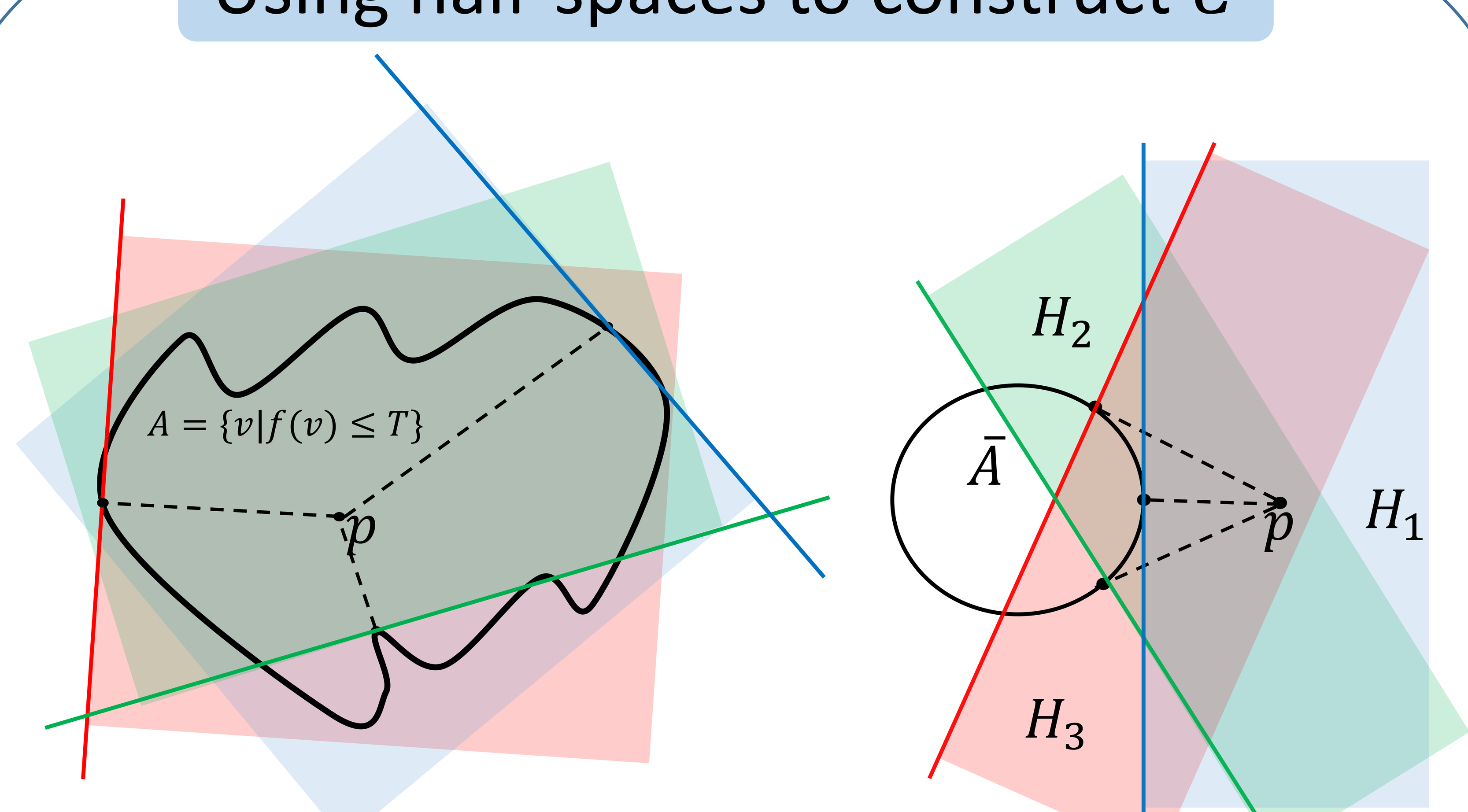


Geometric Monitoring

The set $A = \{v | f(v) \le T\}$ Is called the *admissible region*. C is a *convex* subset of A. C's convexity guarantees that if $v_1, v_2 \dots v_k \in C$, then $v = \left(\frac{\sum v_i}{k}\right) \in C \subseteq A$. Node i stays silent as long as $v_i \in C$.

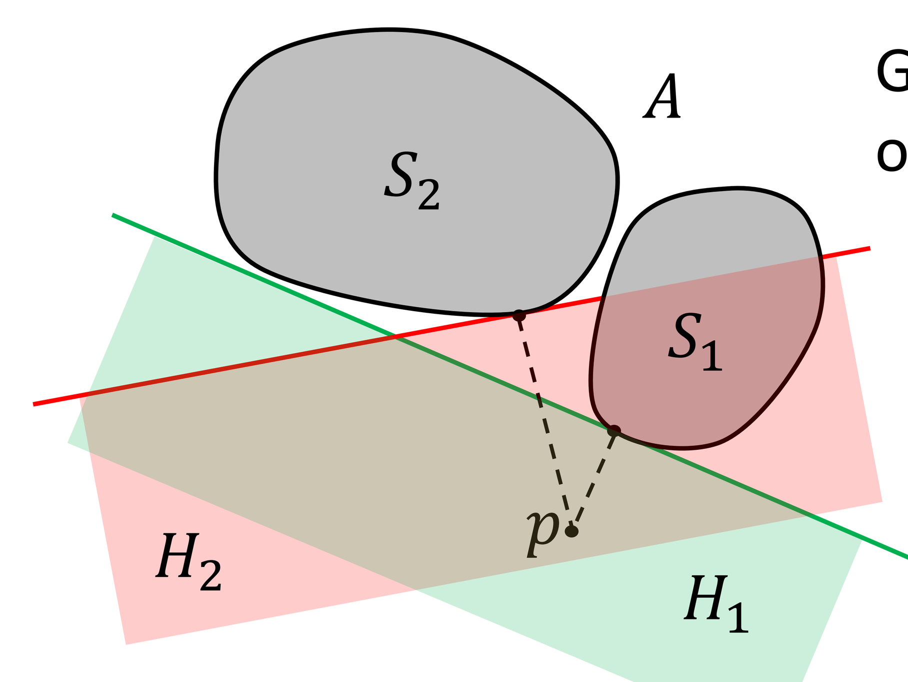


Using half-spaces to construct C



Left: A convex subset of A containing a point p can be created in the following manner: for each point q on the boundary of $\bar{A}$ take the half-space orthogonal to pq containing p. Then take the intersection of these half-spaces. Right: When $\bar{A}$ is convex this method yields a non-optimal C. Clearly, H_1 is a better convex subset (it contains $H_1 \cap H_2 \cap H_3$).

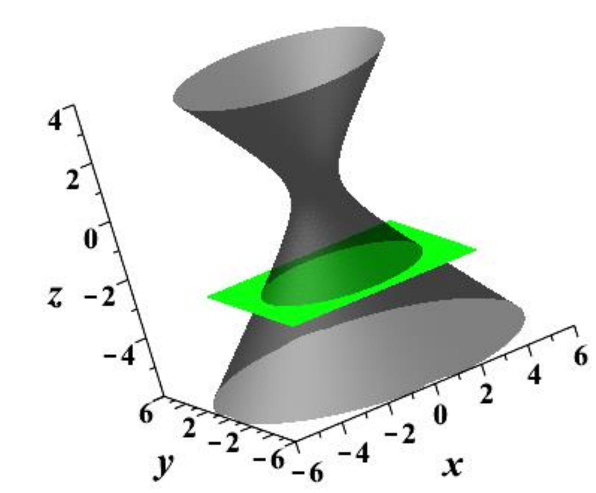
Convex Decomposition



General case: *convex decomposition* of $\bar{A}$, but avoid ***redundancy!***

If $\bar{A}$ is the union of the two convex sets, S_1 and S_2, and H_1 separates p from S_2, there is no need to intersect with H_2, and C can be defined as H_1

If $\bar{A}$ is the hyperboloid, it can be decomposed to a infinite union of (convex) parallel discs.



Monitoring the median

- We wish to monitor $med(x_1, x_2, x_3) > 0$
- A is not convex. The vectors $(1,1,-11)$, $(1,-11,1)$, $(-11,1,1)$ are in A, but their average $(-3,-3,-3)$ is not.
- We partition the complement $\bar{A}$ to convex subsets: $(x_1, x_2 \le 0)$, $(x_1, x_3 \le 0)$, $(x_2, x_3 \le 0)$
- This is a non-redundant decomposition (can be proved)
- It can be generalized to any dimension.

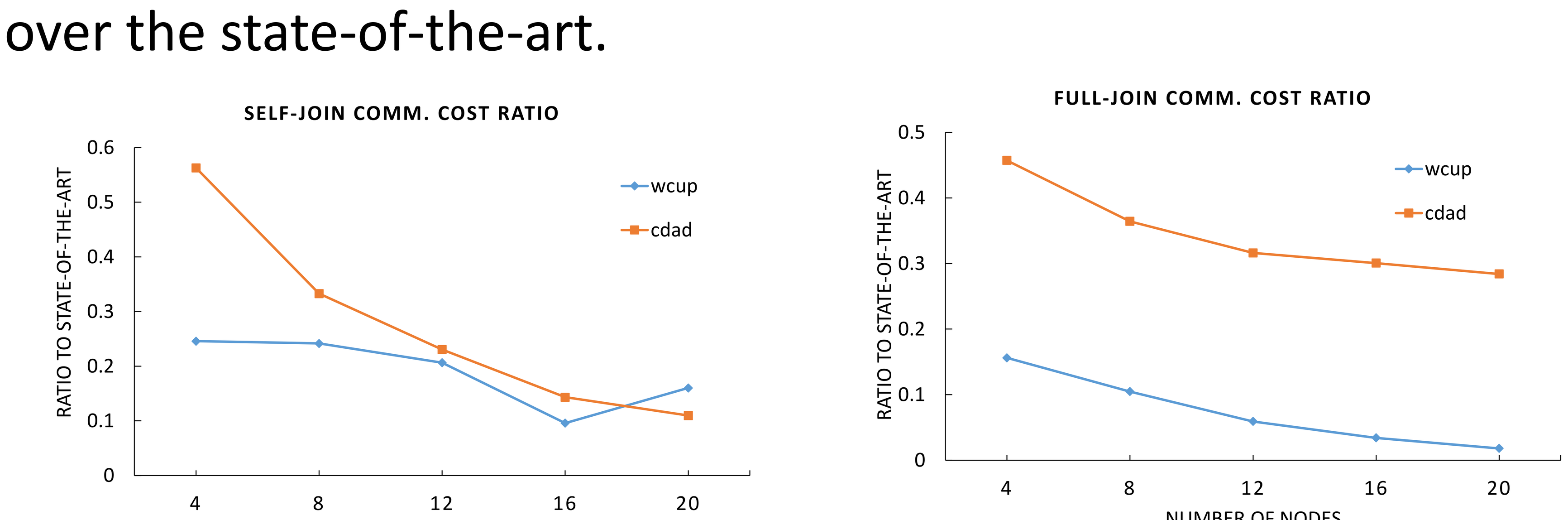
* The number of sets in the decomposition is exponential, but a simple trick allows to check whether a point belongs to the A in $n\log(n)$ time, where n is the dimension.

Sketches

- AGMS sketches provide summary of the data using a $k \times l$ matrix M.
- To approximate the result of range, self-join and full-join queries the median operator is applied to a vector constructed either by taking the norm squared of each matrix row, or by taking the inner-product of the rows of two sketch matrices (median of a *quadric* function).
- This is more complicated than the linear median case, but can be solved.

Evaluation

The CD method was tested on two real data sets for self-join and full-join queries using sketches, it yielded substantial improvement over the state-of-the-art.



Communication cost (bytes sent) relative to state-of-the-art for varying number of nodes over two datasets. Left: self-join. Right: full-join (smaller is better).