Querying Big, Dynamic, Distributed Data

Minos Garofalakis

Technical University of Crete

Software Technology and Network Applications Lab

LIFT Cast: **Antonios Deligiannakis, Vasilis Samoladas, Odysseas Papapetrou, Nikos Giatrakos (TUC); Daniel Keren (Haifa U), Assaf Schuster, Tsachi Sharfman (Technion)**

Big Data is Big News (and Big Business…)

- Rapid growth due to several information-generating technologies, such as mobile computing, sensornets, and social networks

- How can we cost-effectively manage and analyze all this data…?

The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA
InformationWeek
BIG DATA
You have terabytes aplenty. Now you need speed: faster queries, real-time insights. Here's how to get there. p.22
By Doug Henschen

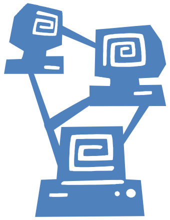
Big Data Challenges: The Four V's (and one D)…

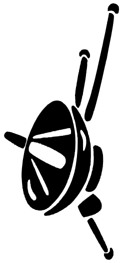
- **Volume:** Scaling from Terabytes to Exa/Zettabytes
- **Velocity:** Processing massive amounts of ***streaming data***
- **Variety:** Managing the complexity of multiple relational and non-relational data types and schemas
- **Veracity:** Handling the inherent uncertainty and noise in the data
- **Distribution:** Dealing with **massively distributed** information
- ***LIFT focus: Volume, Velocity, Distribution***

Velocity: *Continuous Stream Querying*

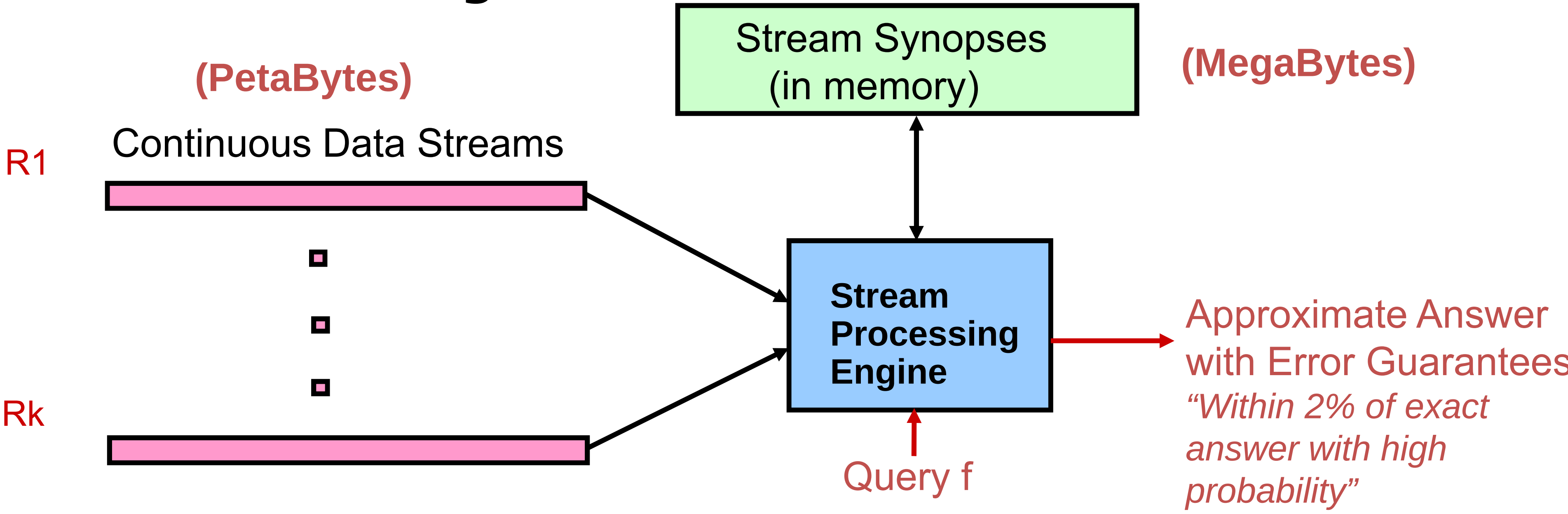
There are many scenarios where we need to monitor/track events over streaming data:

- Network health monitoring within a large ISP
- Collecting and monitoring environmental data with sensors
- Observing usage and abuse of large-scale data centers





Stream Processing Model



Stream Synopses (in memory)

(MegaBytes)

(PetaBytes)

Continuous Data Streams

R1

Rk

Stream Processing Engine

Query f

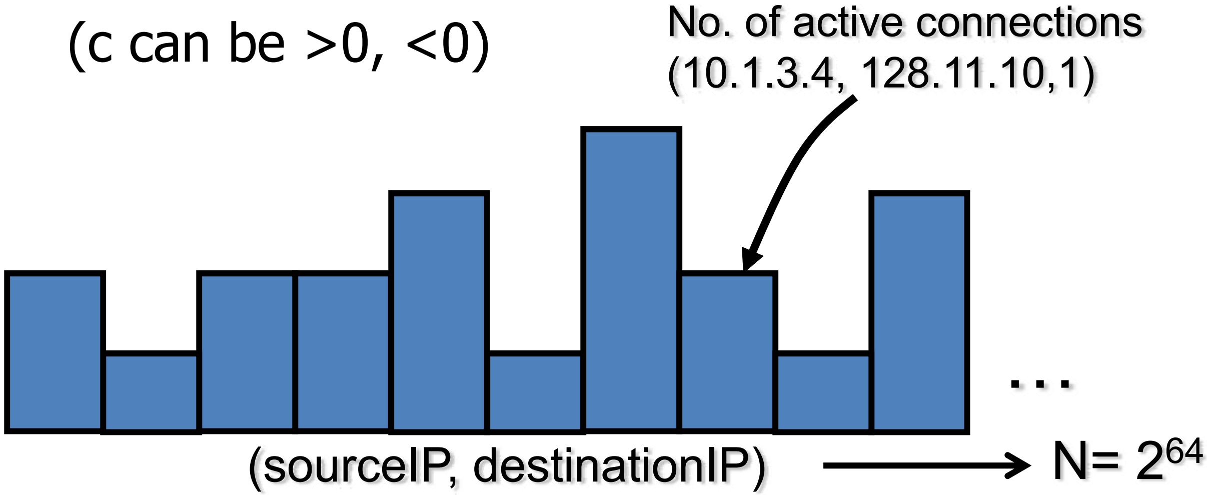
Approximate Answer with Error Guarantees

"Within 2% of exact answer with high probability"

- Approximate answers often suffice, e.g., trends, anomalies
- Requirements for stream synopses
  - *Single Pass:* Each record is examined at most once, in arrival order
  - *Small Space:* Log or polylog in data stream size
  - *Small Time:* Per-record processing time must be low
  - Also: *Delete-proof, Composable*, …

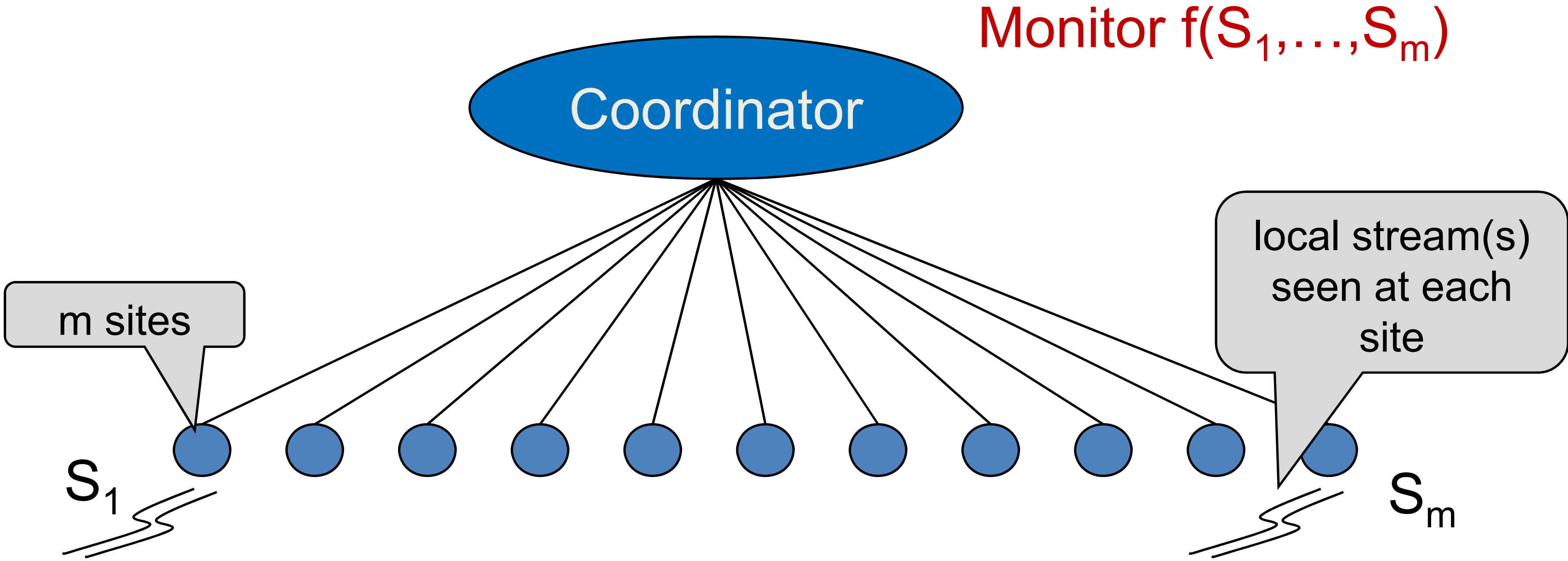
Model of a Relational Stream

- Relation "signal": *Large* array $v_S[1…N]$ with values $v_S[i]$ initially zero
  - Frequency-distribution array of **S**
  - Multi-dimensional arrays as well (e.g., row-major)
- **Relation implicitly rendered via a *stream of updates***
  - Update <x, c> implying
    - $v_S[x] := v_S[x] + c$ (c can be >0, <0)



- ***Goal:* Compute queries (functions) on such dynamic vectors in "small" space and time (<< N)**

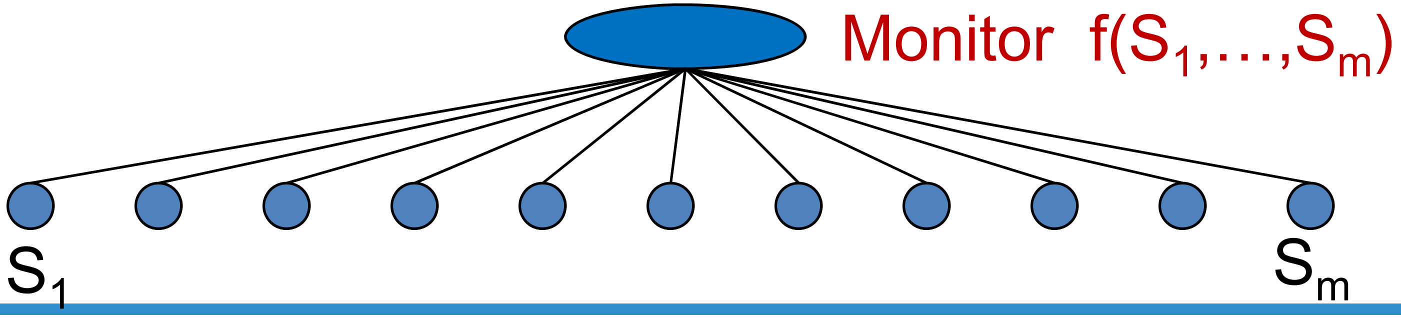
Velocity & Distribution: *Continuous Distributed Streaming*



- Other structures possible (e.g., hierarchical, P2P)
- Goal: *Continuously track* (global) query over streams at the coordinator
  - Using small space, time, and ***communication***
  - Example queries: Join aggregates, Variance, Entropy, Information Gain, …

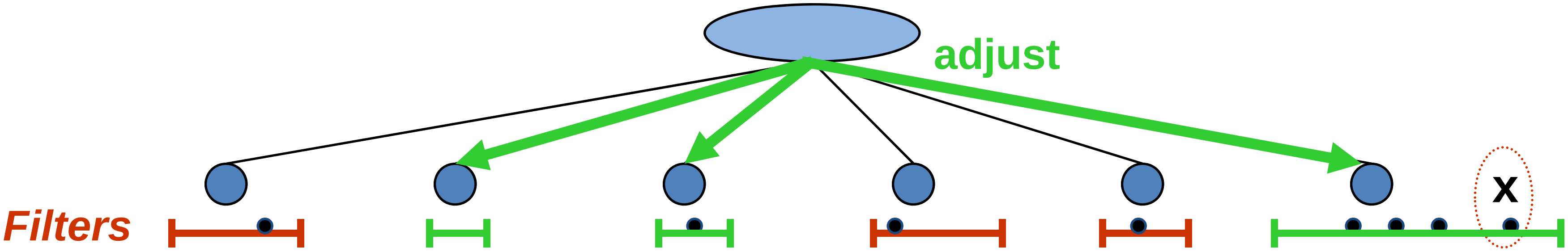
Continuous Distributed Streaming

- But… local site streams continuously change! New readings/data…
- Classes of monitoring problems
  - **Threshold Crossing**: Identify when $f(S) > \tau$
  - **Approximate Tracking**: $f(S)$ within some **guaranteed accuracy bound ε**
    - Tradeoff *accuracy and communication / processing cost*
- Naïve solutions must *continuously* centralize all data
  - Enormous communication overhead!
- Instead, ***in-situ*** stream processing using ***local constraints*** !



Communication-Efficient Monitoring

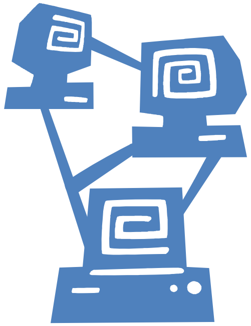
- **Key Idea: *"Push-based" in-situ processing***
  - *Local filters* installed at sites process local streaming updates
    - Offer bounds on local-stream behavior (at coordinator)
  - *"Push"* information to coordinator only when filter is violated
  - ***"Safe"!*** Coordinator sets/adjusts local filters to guarantee accuracy



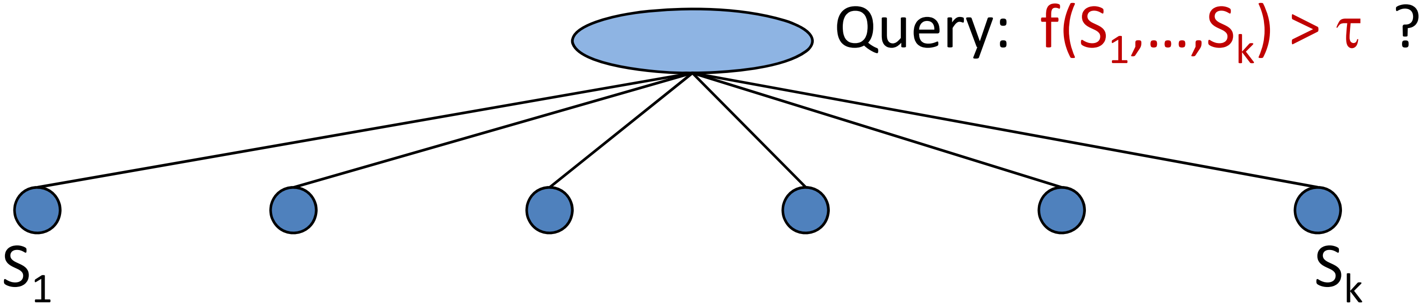
- Easy for linear functions! Exploit additivity…
- *Non-linear f() …??*

Outline

- Introduction: Continuous Distributed Streaming
- The Geometric Method (GM)
- Recent Work: GM + Sketches
- Challenges & Conclusion



Monitoring General, Non-linear Functions



- For general, non-linear f(), the problem becomes a lot harder!
  - E.g., information gain over global data distribution
- Non-trivial to decompose the global threshold into “safe” local site constraints
    - E.g., consider $N=(N_1+N_2)/2$ and $f(N) = 6N - N^2 > 1$
    Tricky to break into thresholds for $f(N_1)$ and $f(N_2)$

The Geometric Method

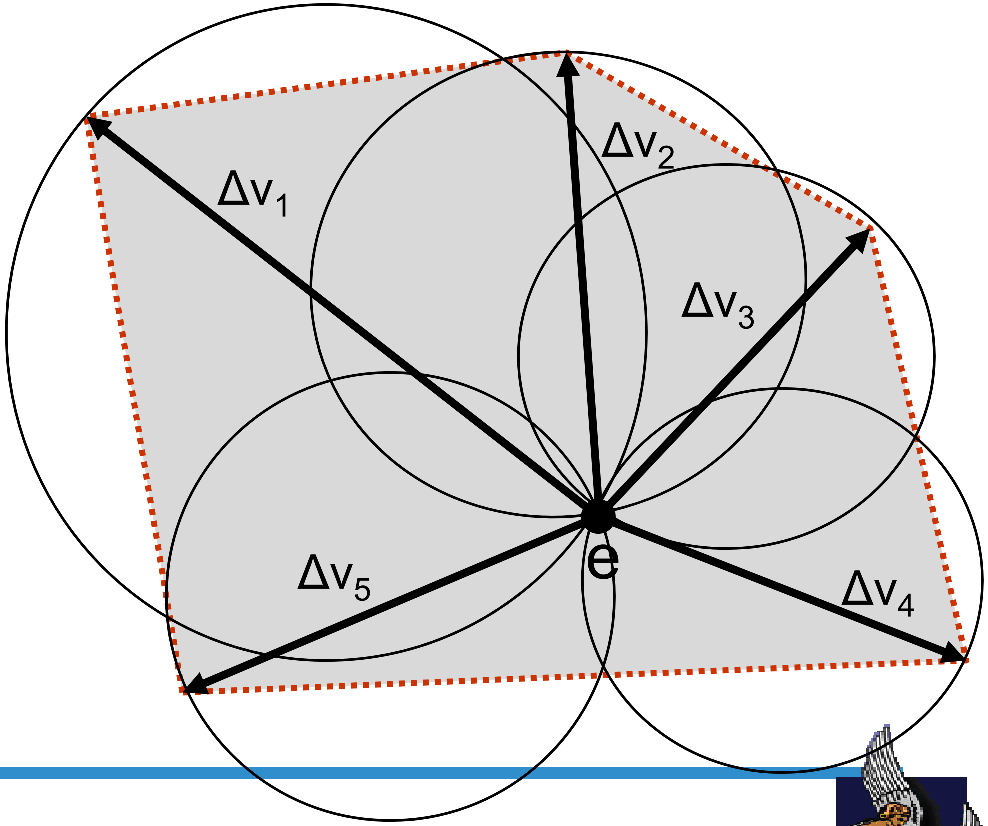
- A general purpose geometric approach [SKS SIGMOD’06]
  - Monitor **function domain** rather than the range of values!
- Each site tracks a local statistics *vector* v_i (e.g., data distribution)
- Global condition is $f(v) > \tau$, where $v = \sum_i \lambda_i v_i$ $(\sum_i \lambda_i = 1)$
  - v = convex combination of local statistics vectors
- All sites share estimate $e = \sum_i \lambda_i v_i'$ of v

  based on latest update v_i' from site i
- Each site i tracks its drift from its most recent update $\Delta v_i = v_i - v_i'$

Covering the Convex Hull

- **Key observation:** $\mathbf{v = \sum_i \lambda_i \cdot (e + \Delta v_i)}$
  (a convex combination of "translated" local drifts)

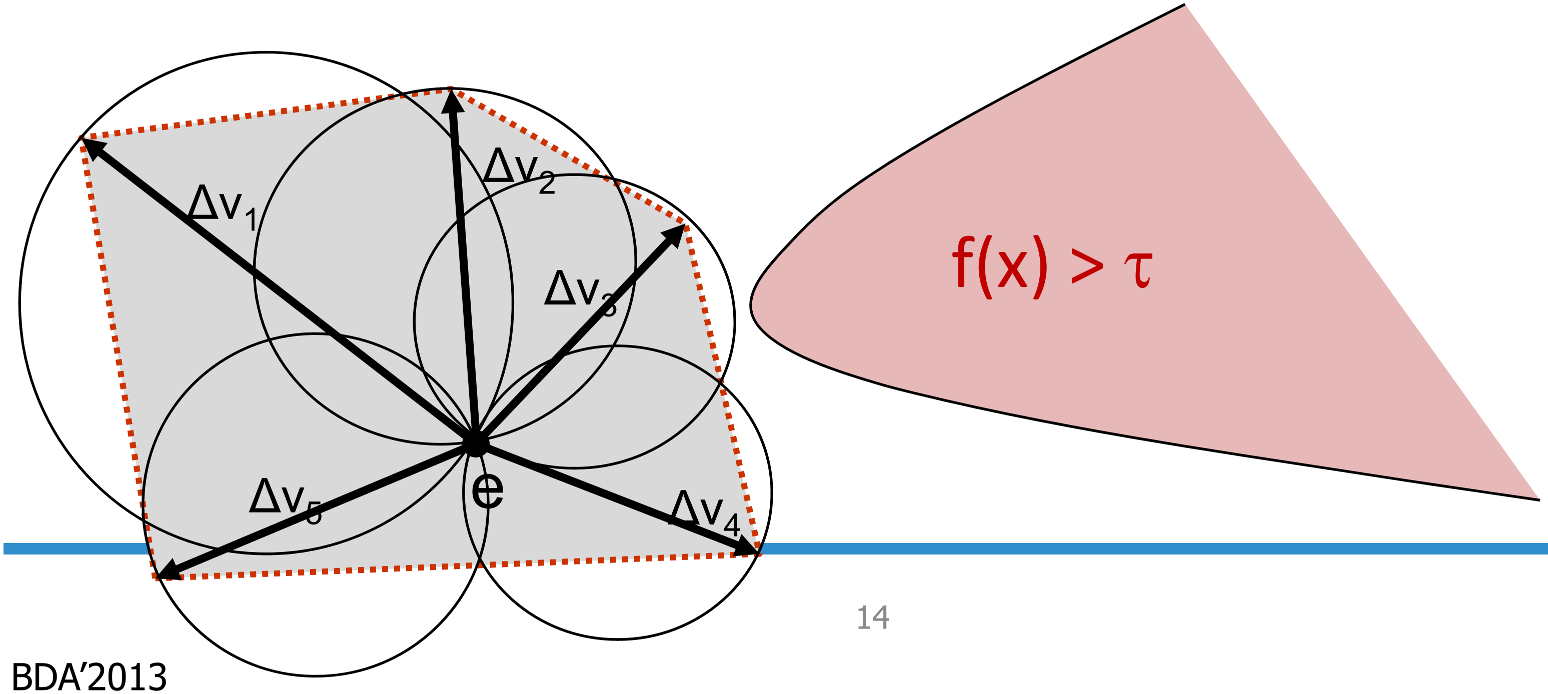
- v lies in the convex hull of the $(e+\Delta v_i)$ vectors
- Convex hull is completely covered by spheres with radii $||\Delta v_i/2||_2$ centered at $e+\Delta v_i/2$
- Each such sphere can be constructed independently



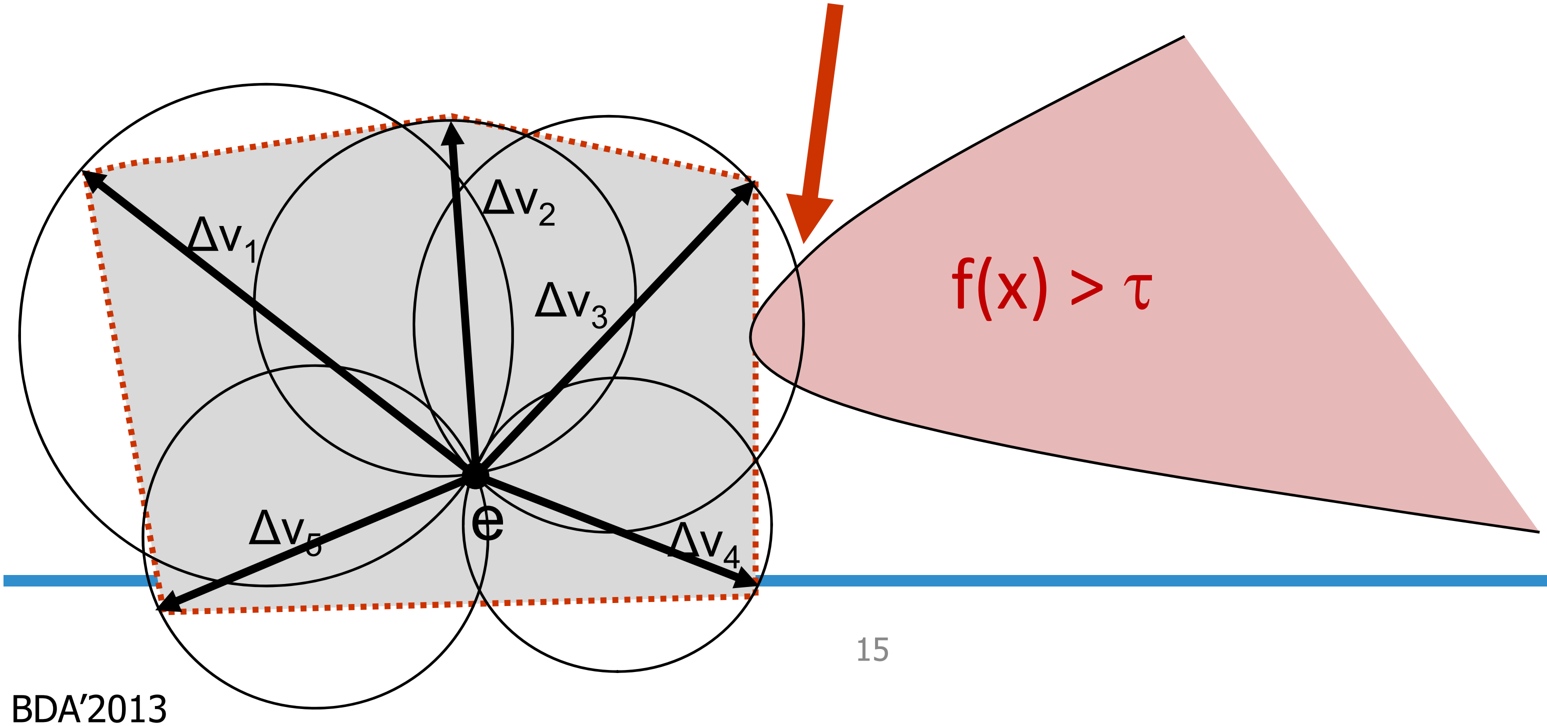
Monochromatic Regions

- Monochromatic Region: For all points x in the region f(x) is on the same side of the threshold (f(x) > τ or f(x) $\leq$ τ)

- Each site independently checks its sphere is monochromatic
  - Find max and min for f() in local sphere region (may be costly)
  - Send updated value of v_i if not monochrome

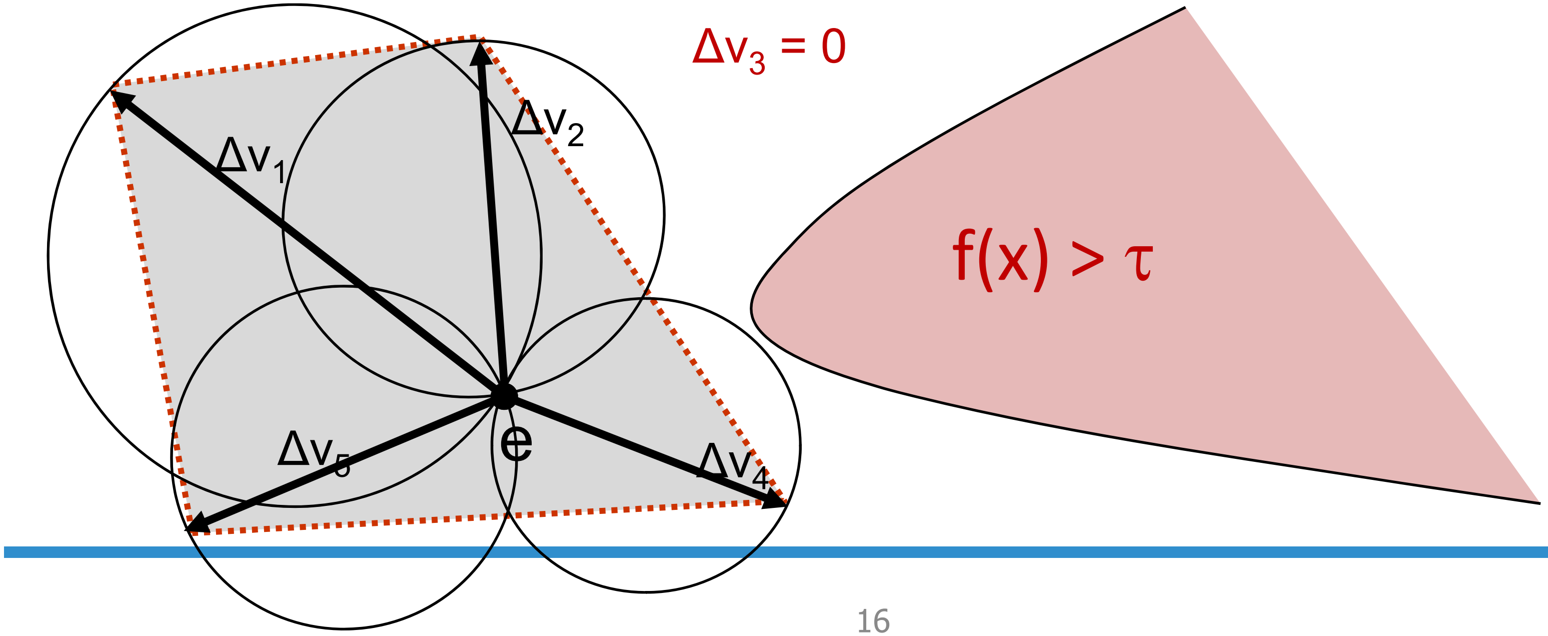


Restoring Monochromicity



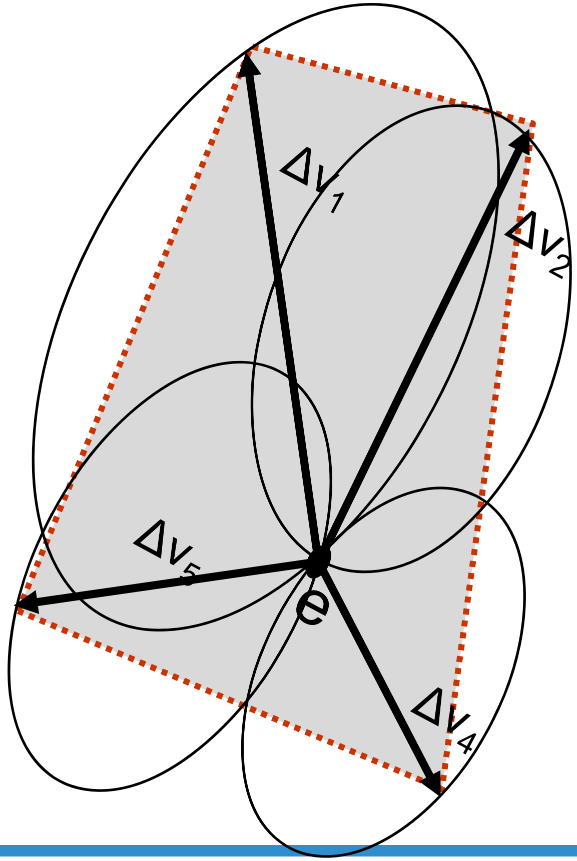
Restoring Monochromicity

- After update, $||\Delta v_i||_2 = 0$ ⇒ Sphere at i is monochromatic
  - Global estimate e is updated, which may cause more site update broadcasts
- Coordinator case*:* Can allocate local slack vectors to sites to enable "localized" resolutions
  - Drift (=radius) depends on slack (adjusted locally for subsets)



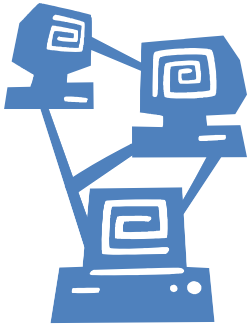
Extensions: Transforms, Shifts, and Safe Zones

- **Subsequent developments** [SKS TKDE'12]
  - Same analysis of correctness holds when spheres are allowed to be ellipsoids
  - Different reference vectors can be used to increase radius when close to threshold values
  - Combining these observations allows additional cost savings
- **More general theory of "Safe Zones"**
  - Convex subsets of the admissible region

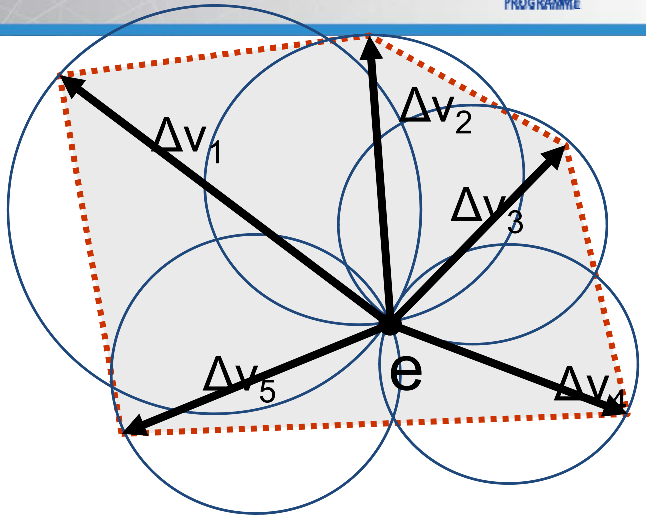


Outline

- Introduction: Continuous Distributed Streaming
- The Geometric Method (GM)
- Recent Work: GM + Sketches
- Challenges & Conclusion

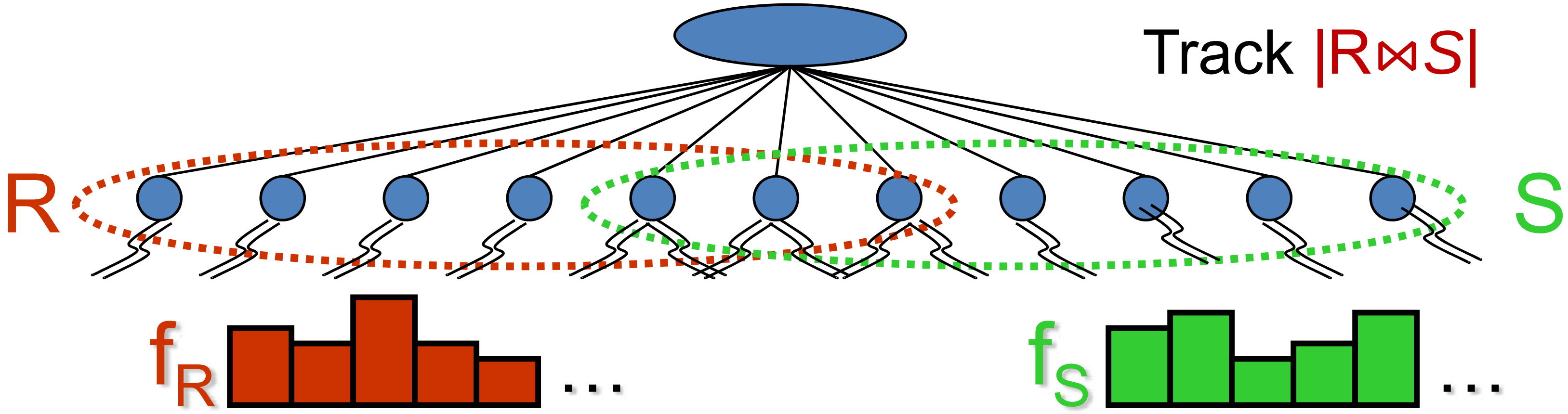


Geometric Query Tracking using AMS Sketches [GKS VLDB'13]



- ***Continuous approximate monitoring*** rather than simple threshold crossing
  - Maintain the value of a function to within specified accuracy bound ε
- Too much local information ➔ *Local summaries at sites*
  - A form of dimensionality reduction
  - Bounding regions for the *lower-dimensional sketching-space domain*
  - Function over sketch => Sketching error θ
    - Accounted for in the region checks (depend on both ε, θ)
- ***Key Problems: (1) Minimize data exchange volume (2) Deal with highly-nonlinear AMS estimator***

Tracking Complex Aggregate Queries

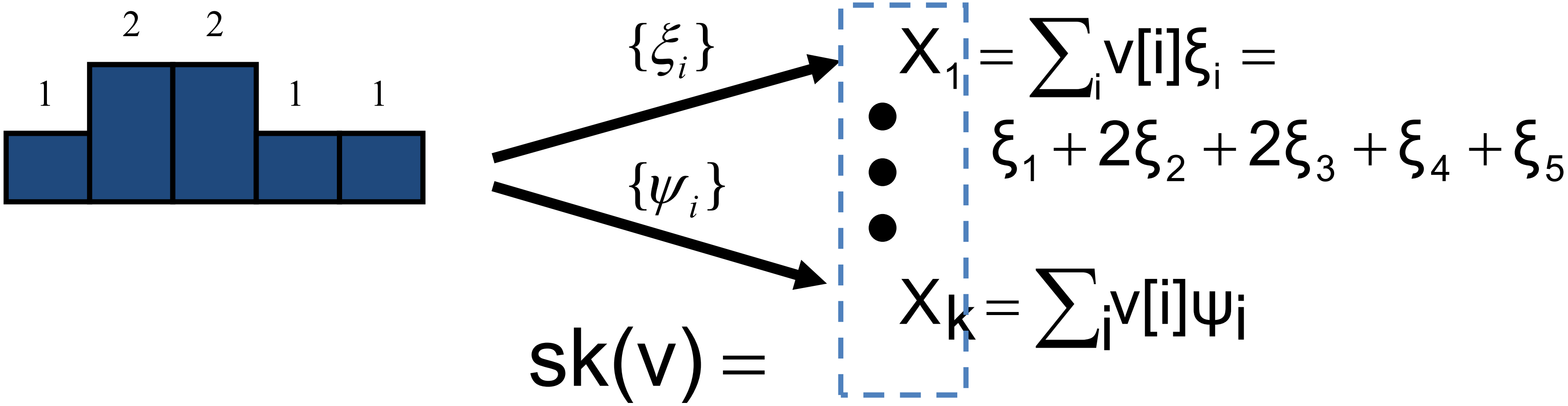


- *Class of queries:* Generalized inner products of streams

$$|R \bowtie S| = f_R \cdot f_S = \sum_v f_R[v]\, f_S[v]$$

  - Join/multi-join aggregates, range queries, heavy hitters, histograms, wavelets, …

AMS Sketches 101



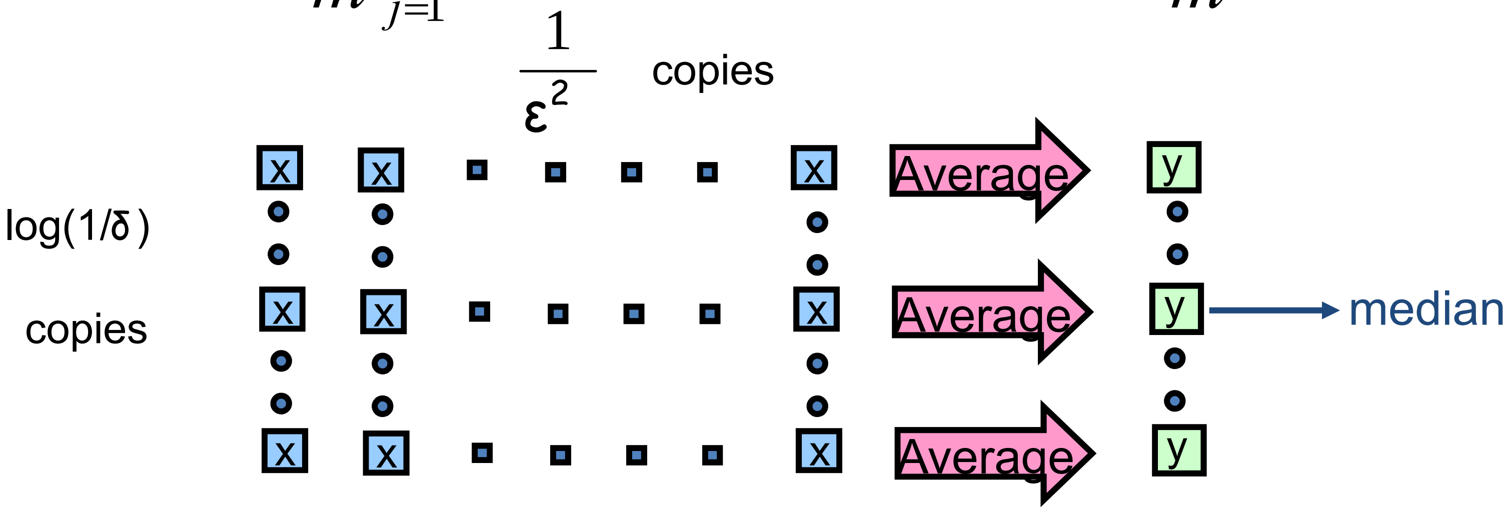
$$sk(v) = \begin{array}{l} X_1 = \sum_i v[i]\xi_i = \xi_1 + 2\xi_2 + 2\xi_3 + \xi_4 + \xi_5 \\ \bullet \\ \bullet \\ \bullet \\ X_k = \sum_i v[i]\psi_i \end{array}$$

- Simple randomized linear projections of data distribution
  - Easily computed over stream using logarithmic space
  - *Linear:* Compose through simple vector addition

Monitored Function…?

AMS Estimator function **for Self-Join**

$$f(sk(v)) = median_{i=1..n}\{\frac{1}{m}\sum_{j=1}^{m} sk(v)[i,j]^2\} = median_{i=1..n}\{\frac{1}{m}\|sk(v)[i]\|^2\}$$



- Theorem (AMS96): Sketching approximates $\|v\|_2^2$ to within an error of $\pm\varepsilon\|v\|_2^2$ with probability $\geq 1-\delta$ using $O(\frac{1}{\varepsilon^2}\log(1/\delta))$ counters

Geometric Function Monitoring using AMS Sketches [GKS VLDB'13]

- **Sketches can still get pretty large!**

- **Minimizing volume of data exchanges**
  - Can reduce problem to monitoring in O(log(1/δ)) dimensions
  - Local Stats vector: Row-norm error-vector **d** defined as

$$d[i] = \| sk(v)[i] - sk(v')[i] \|$$

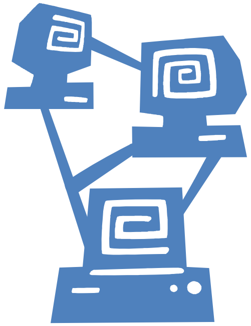
  - Using triangle inequality and median monotonicity, can bound the AMS estimator using functions of **d**
  - GM monitoring of **f(d)** -- only O(log(1/δ)) dimensions!

Geometric Function Monitoring using AMS Sketches [GKS VLDB'13]

- **Efficiently deciding ball monochromicity for the median operator**
  - Fast greedy algorithm for determining the distance to the inadmissible region
- *(Non-trivial!)* extension to *general inner product (join) queries*
- Consistent communication cost gains 30-40% over earlier sketch-based methods; Over 100% in terms of sketch-data exchanges!

Outline

- Introduction: Continuous Distributed Streaming
- The Geometric Method (GM)
- Recent Work: GM + Sketches
- Challenges & Conclusion

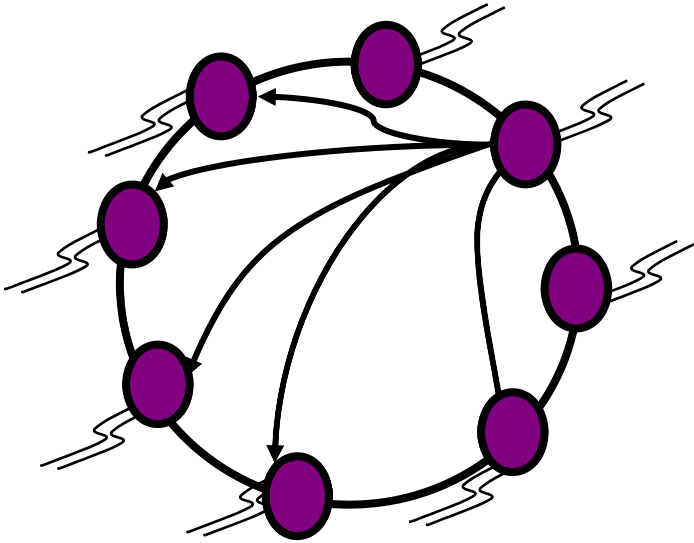


Work in CD Monitoring

- Much interest in these problems in TCS and Database areas
- Many specific functions of (global) data distribution studied:
  - Set expressions [Das,Ganguly,G,Rastogi'04]
  - Quantiles and heavy hitters [Cormode,G, Muthukrishnan, Rastogi'05]
  - Number of distinct elements [Cormode et al.,'06]
  - Spectral properties of data matrix [Huang,G, et al.'06]
  - Anomaly detection in networks [Huang ,G, et al.'07]
  - Samples [Cormode et al.'10]
  - Counts, frequencies, ranks [Yi et al.,'12]
- See proceedings of recent NII Shonan meeting on Large-Scale Distributed Computation

http://www.nii.ac.jp/shonan/seminar011/

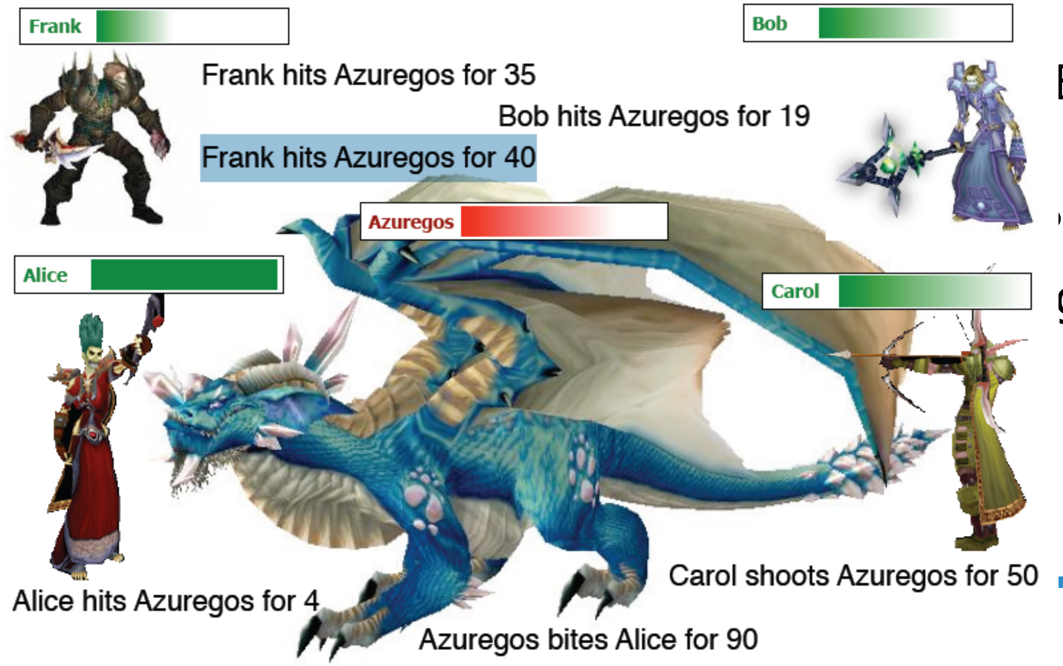
CD Monitoring in Scalable Network Architectures



- E.g., DHT-based P2P networks
- Single query point
  - "Unfolding" the network gives a hierarchy
  - But, single point of failure (i.e., root)
- Decentralized monitoring
  - Everyone participates in computation, all get the result
  - Exploit epidemics? Latency might be problematic…

Monitoring Systems

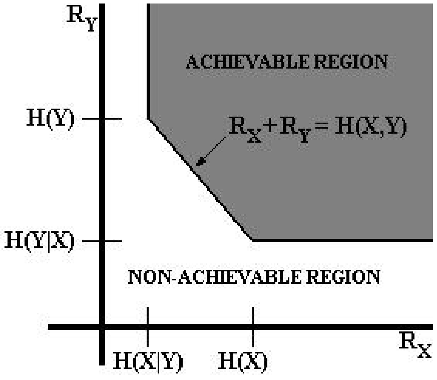
- **Much theory developed, but less progress on deployment**
- **Some empirical study in the lab, with recorded data**
- **Still applications abound: Online Games** [Heffner, Malecha'09]
  - Need to monitor many varying stats and bound communication
- **Several steps to follow:**
  - Build li
  - Evolve BMSs?)
- **Several (**
  - What f
  - What k g?

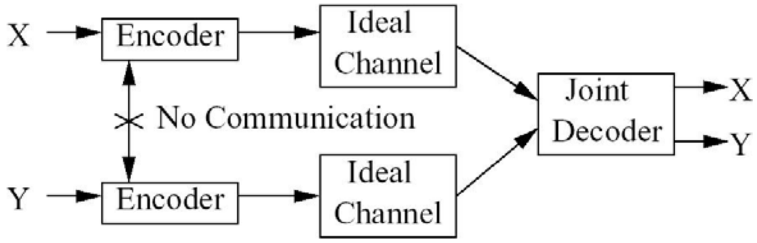

Frank
Bob
Frank hits Azuregos for 35
Bob hits Azuregos for 19
Frank hits Azuregos for 40
Azuregos
Alice
Carol
Carol shoots Azuregos for 50
Alice hits Azuregos for 4
Azuregos bites Alice for 90

Theoretical Foundations

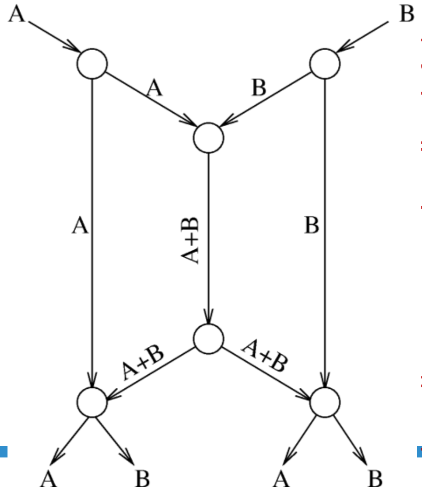
"Communication complexity" studies lower bounds of distributed one-shot computations

- **Gives lower bounds for various problems, e.g., `count distinct` (via reduction to abstract problems)**
- **Need new theory for continuous computations**
  - Based on info. theory and models of how streams evolve?
  - Link to distributed source coding or network coding?





Slepian-Wolf theorem [Slepian Wolf 1973]



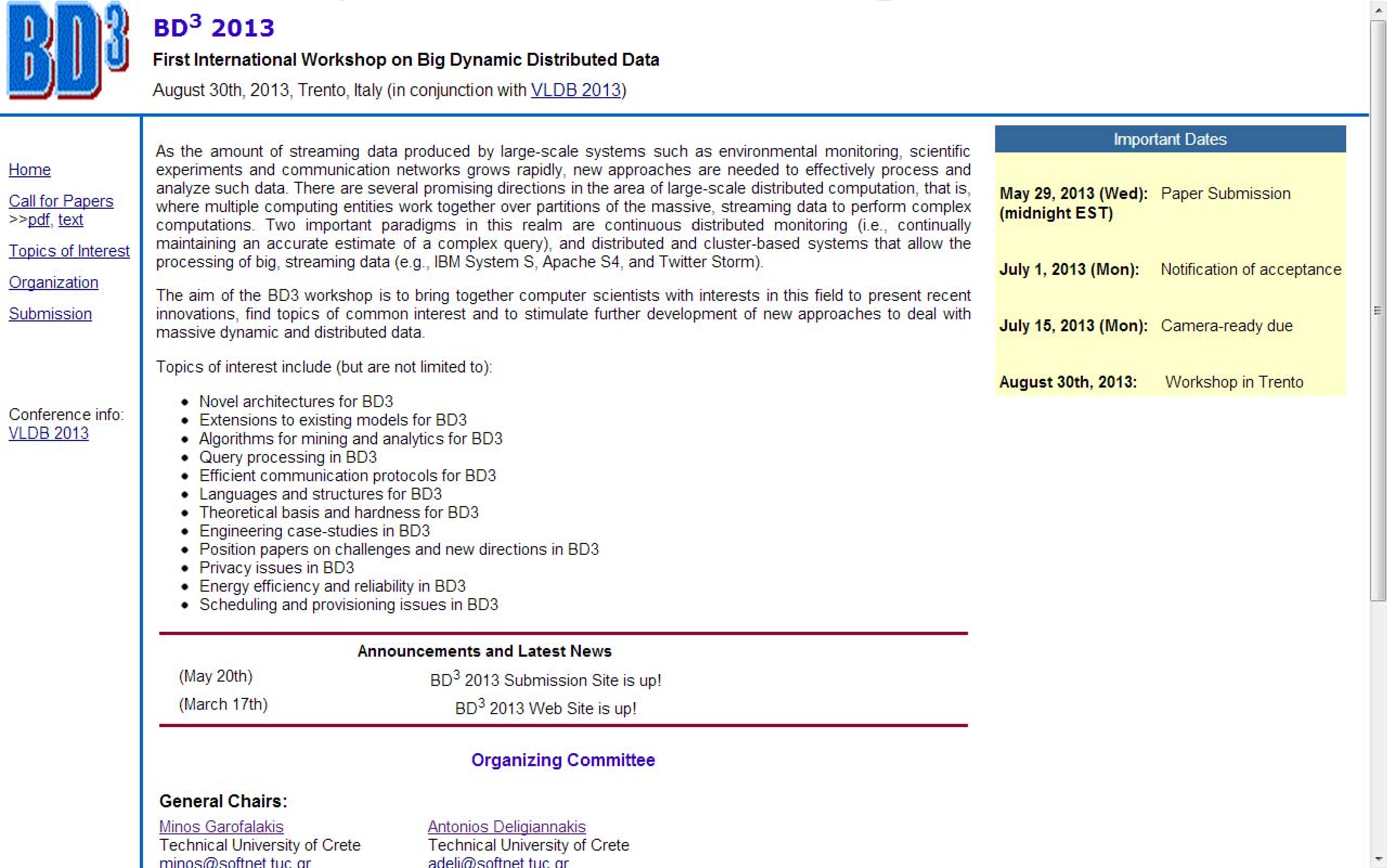
http://www.networkcoding.info/

https://buffy.eecs.berkeley.edu/PHP/resabs/resabs.php?f_year=2005&f_submit=chapgrp&f_chapter=1

Conclusions

- Continuous querying of distributed streams is a natural model
  - Interesting space/time/communication tradeoffs
  - Captures several real-world applications
- **Geometric Method** : Generic tool for monitoring complex, non-linear queries
  - Sketches, dynamic prediction models [GDG SIGMOD'12], recent work on Skyline Monitoring [GP'13]
- Much non-trivial algorithmic and theoretical work in CDS model
  - Intense research interest from DB and TCS communities
  - Deployment in real systems to come…
- ***Much interesting work remains to be done!***

http://www.softnet.tuc.gr/bd3/



BD^3 2013

First International Workshop on Big Dynamic Distributed Data

August 30th, 2013, Trento, Italy (in conjunction with VLDB 2013)

Home

Call for Papers
>>pdf, text

Topics of Interest

Organization

Submission

Conference info:
VLDB 2013

As the amount of streaming data produced by large-scale systems such as environmental monitoring, scientific experiments and communication networks grows rapidly, new approaches are needed to effectively process and analyze such data. There are several promising directions in the area of large-scale distributed computation, that is, where multiple computing entities work together over partitions of the massive, streaming data to perform complex computations. Two important paradigms in this realm are continuous distributed monitoring (i.e., continually maintaining an accurate estimate of a complex query), and distributed and cluster-based systems that allow the processing of big, streaming data (e.g., IBM System S, Apache S4, and Twitter Storm).

The aim of the BD3 workshop is to bring together computer scientists with interests in this field to present recent innovations, find topics of common interest and to stimulate further development of new approaches to deal with massive dynamic and distributed data.

Topics of interest include (but are not limited to):

- Novel architectures for BD3
- Extensions to existing models for BD3
- Algorithms for mining and analytics for BD3
- Query processing in BD3
- Efficient communication protocols for BD3
- Languages and structures for BD3
- Theoretical basis and hardness for BD3
- Engineering case-studies in BD3
- Position papers on challenges and new directions in BD3
- Privacy issues in BD3
- Energy efficiency and reliability in BD3
- Scheduling and provisioning issues in BD3

Important Dates	
May 29, 2013 (Wed): (midnight EST)	Paper Submission
July 1, 2013 (Mon):	Notification of acceptance
July 15, 2013 (Mon):	Camera-ready due
August 30th, 2013:	Workshop in Trento

Announcements and Latest News

(May 20th)	BD^3 2013 Submission Site is up!
(March 17th)	BD^3 2013 Web Site is up!

Organizing Committee

General Chairs:

Minos Garofalakis Technical University of Crete minos@softnet.tuc.gr	Antonios Deligiannakis Technical University of Crete adeli@softnet.tuc.gr

Thank you!

http://www.lift-eu.org/

http://www.softnet.tuc.gr/~minos/