Summarizing 2D Data with Skyline-based Statistical Descriptors

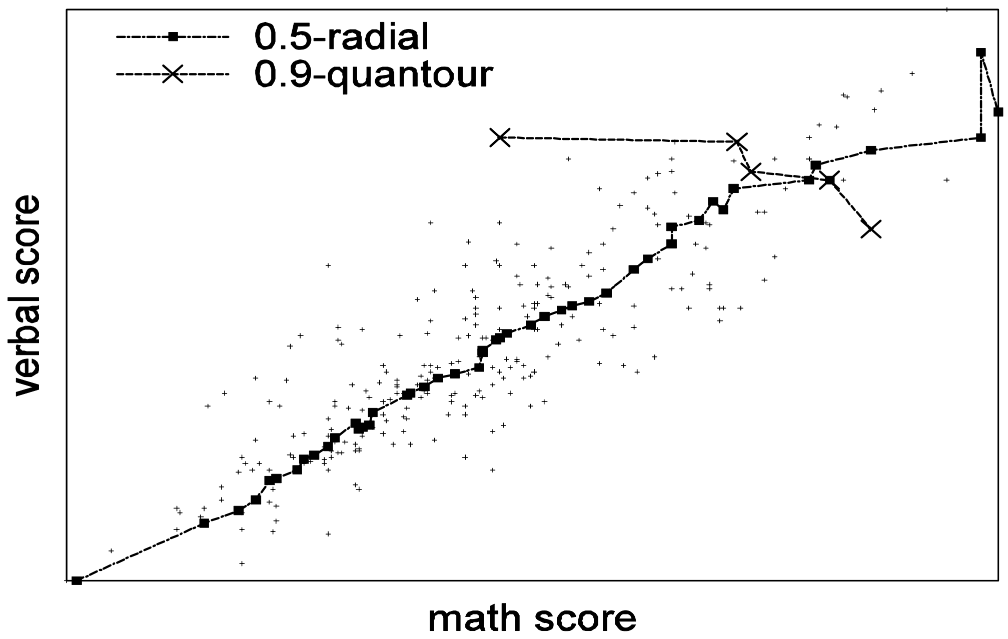
Flip Korn, AT&T Labs

joint work with
Graham Cormode, Divesh Srivastava
AT&T Labs

Printed with FinePrint trial version - purchase at www.fineprint.com

Example: SAT Scores

- Percentiles on each variable not sufficient
- How to compare values across dimensions?
- Dominant students, balanced students



Printed with FinePrint trial version - purchase at www.fineprint.com

Example: Flow Size Distribution

- IP flow summarizes #pkts, #bytes
- Joint distribution indicates behavior
- Track changes in bandwidth, application, etc.
- Detect anomalies (DoS attacks, BGP flaps)

	Many pkts	Few pkts
Many bytes	ftp	http
Few bytes	ping	dhcp

Printed with FinePrint trial version - purchase at www.fineprint.com

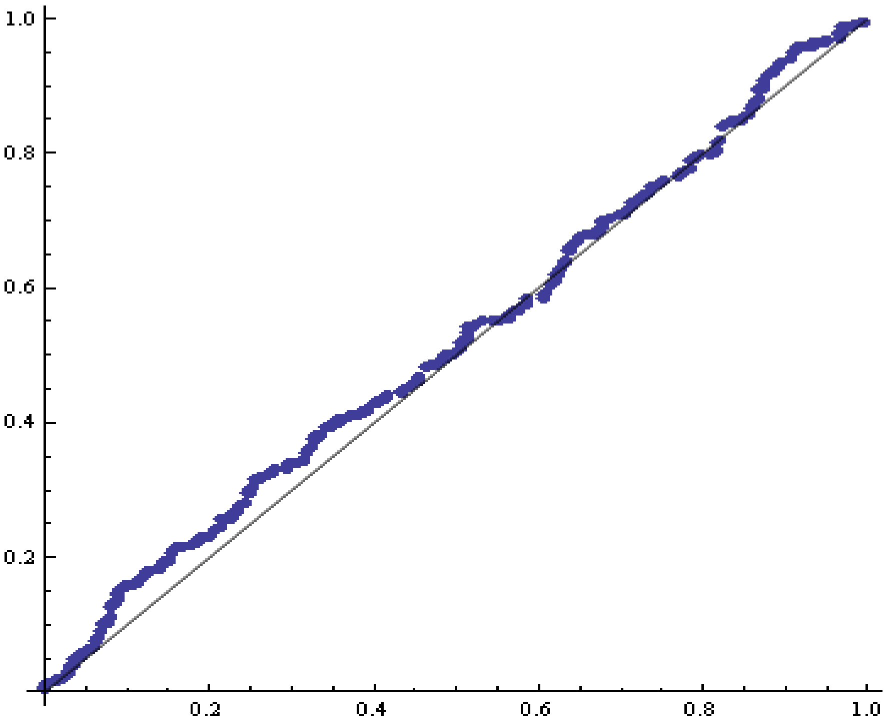
Desiderata

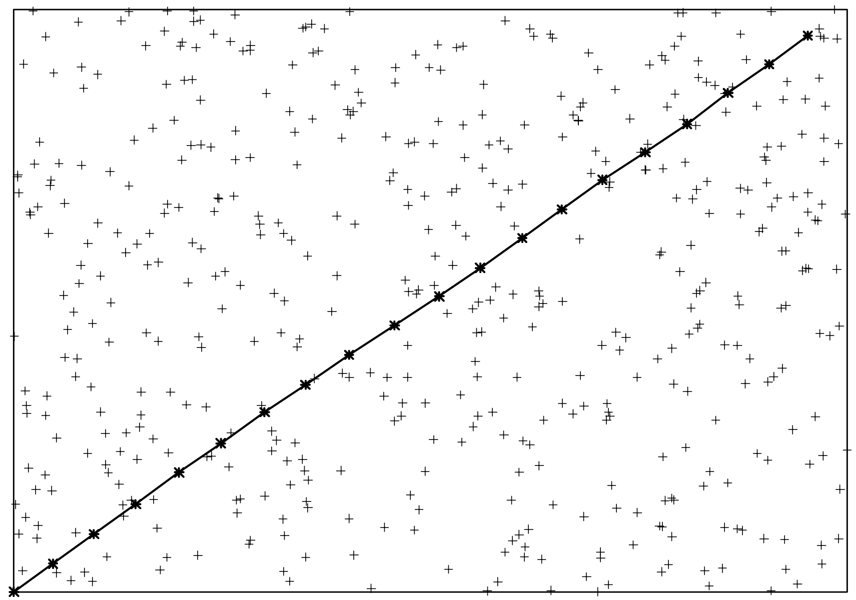
- Quantile: item with rank $\Gamma \phi N$ (*eg*, median)
- But no total (rank) ordering in 2D
- Goal: capture joint distribution
  - Dominance: Pareto optimality
  - Skew: compare trade-offs betw variables
- Robustness is crucial
  - based on *ranks*, not values

Printed with FinePrint trial version - purchase at www.fineprint.com

QQ-Plots

- Quantiles on each dimension
- Doesn't consider joint distribution
- May return points not in data set

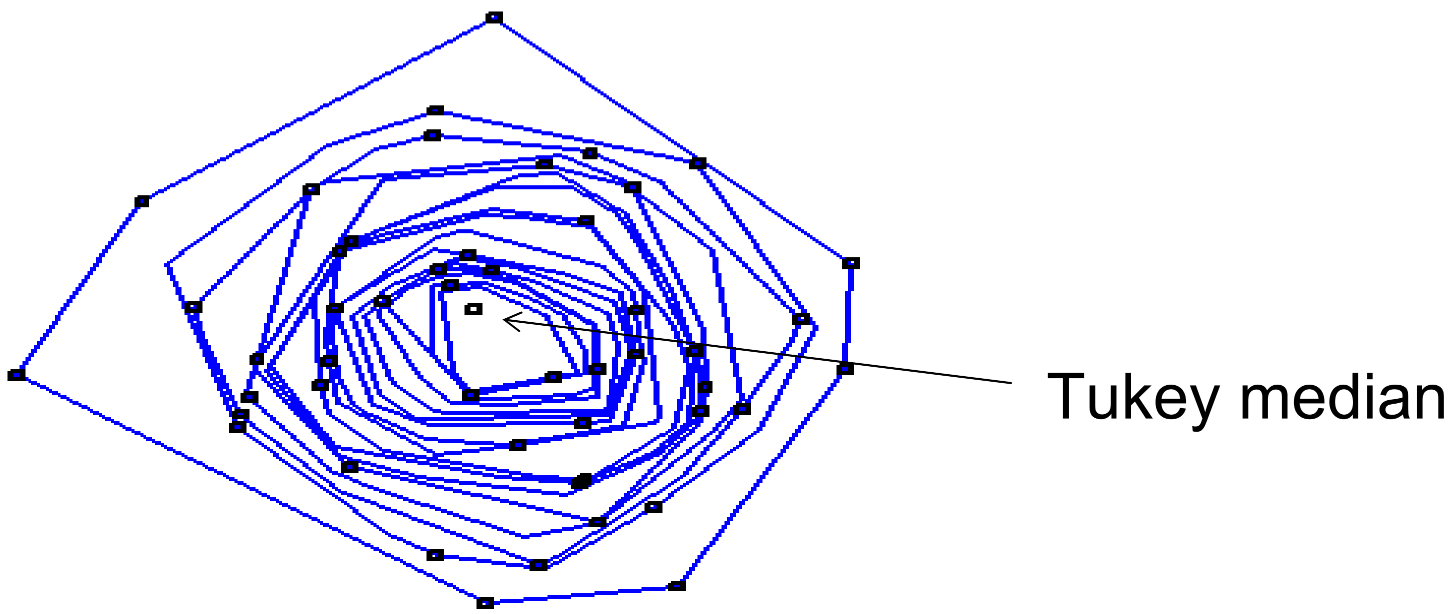




Printed with FinePrint trial version - purchase at www.fineprint.com

Depth Contours

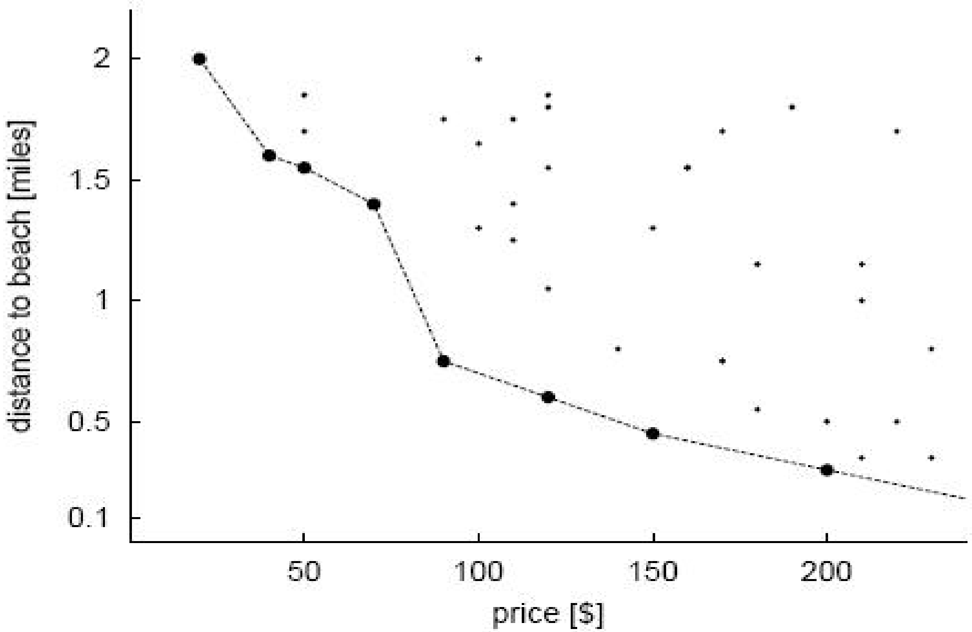
- Recursively compute convex hull
- Dot-product dominance
- Arbitrary #layers, #points per layer



Printed with FinePrint trial version - purchase at www.fineprint.com

Skylines

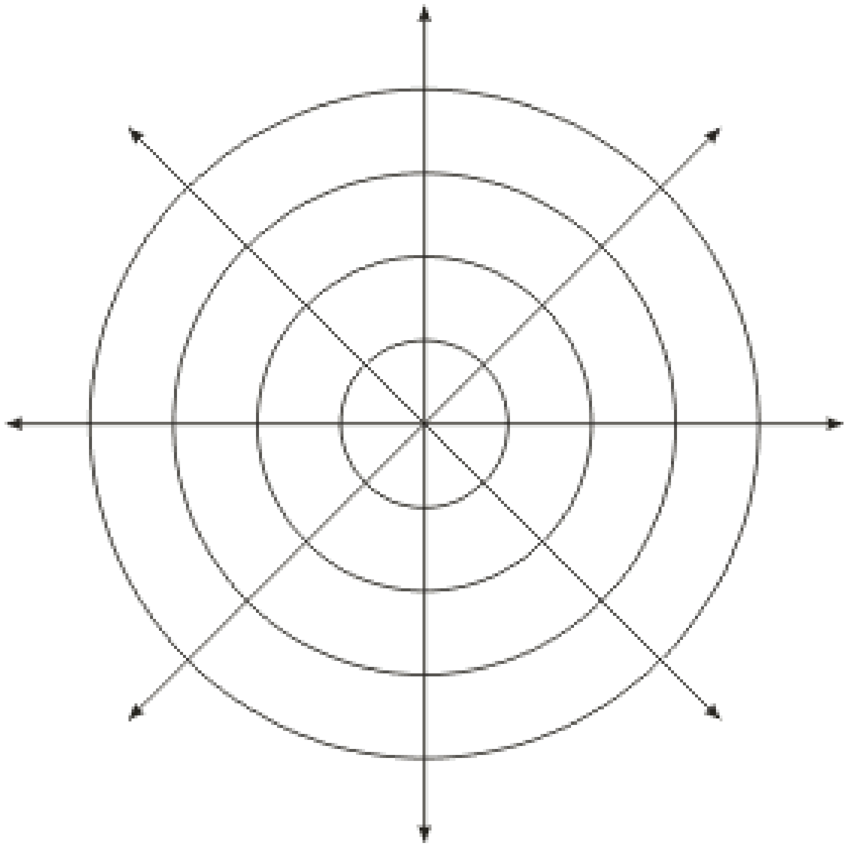
- Points not dominated by a point (*k* points)
- Can be of arbitrary size
- Additional criteria, but not based on distribution



Printed with FinePrint trial version - purchase at www.fineprint.com

Our idea

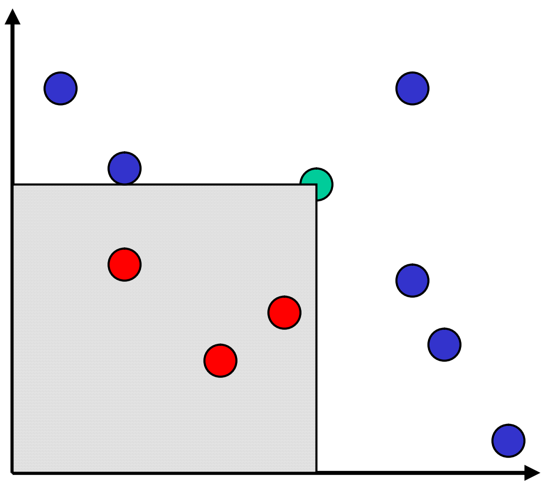
- Simple and natural: inspired by polar coords!

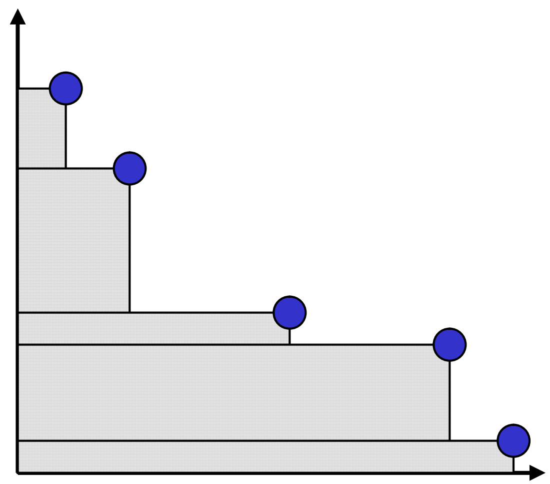


Printed with FinePrint trial version - purchase at www.fineprint.com

ϕ-Quantour Definition

- p dominates q iff ($px \geq qx$) and ($py \geq qy$)
- $P\phi$ = points dominating $\leq \Gamma\phi N$ points
- ϕ-quantour: skyline of $P\phi$

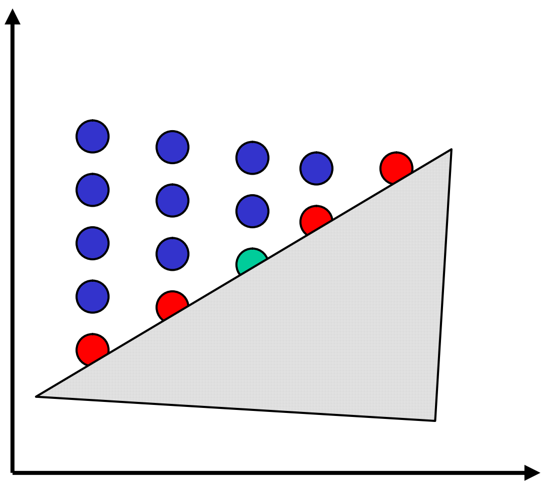


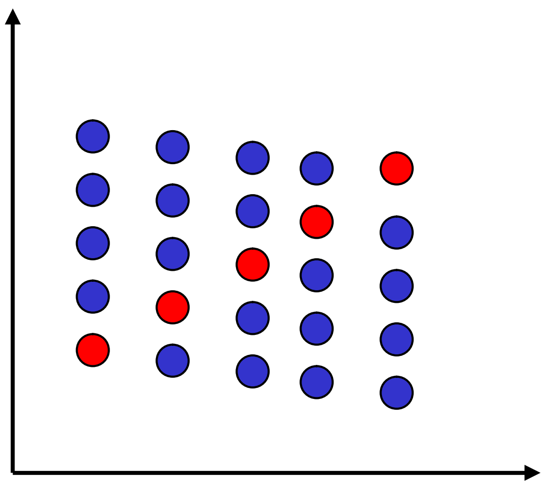


Printed with FinePrint trial version - purchase at www.fineprint.com

α-Radial Definition

- skew = ranky(p) / [xrank(p) + yrank(p)]
- P α = points with skew $\leq \Gamma \alpha N$ points
- α-radial: skyline of P α

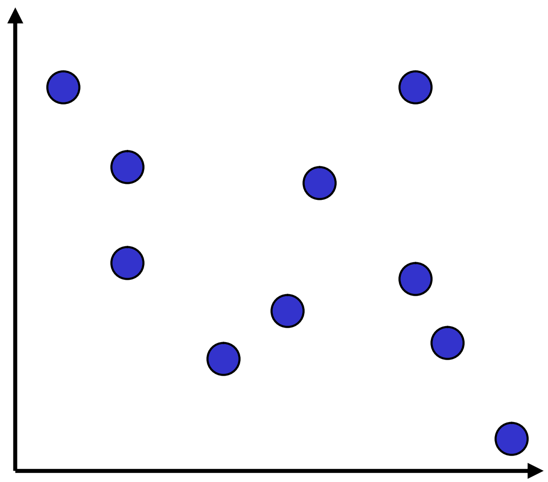




Printed with FinePrint trial version - purchase at www.fineprint.com

(α , ϕ)-Quantile Definition

- ϕ -skyline of α -skyline of P $\phi \cap$ P α

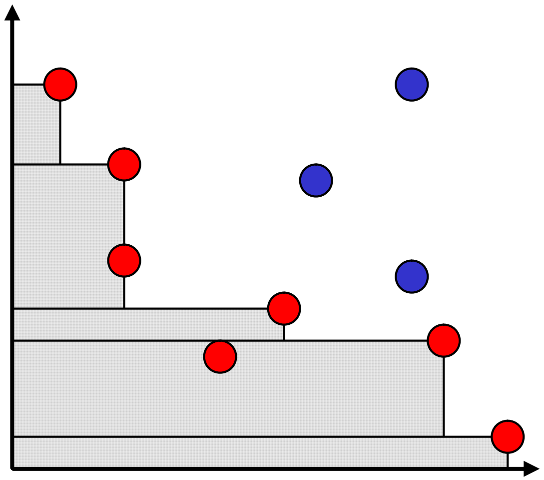


Printed with FinePrint trial version - purchase at www.fineprint.com

(α , ϕ)-Quantile Definition

- ϕ -skyline of α -skyline of P $\phi \cap$ P α

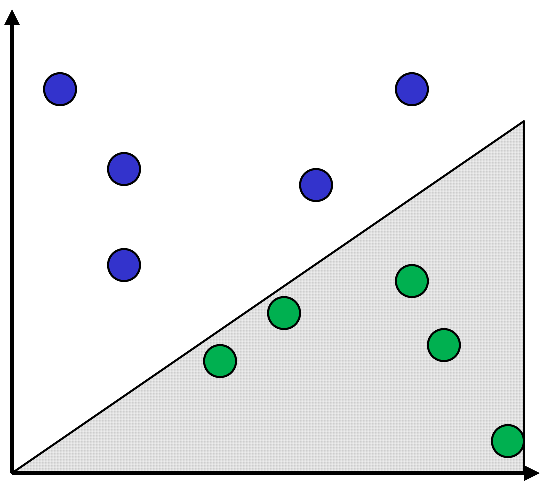
P ϕ



Printed with FinePrint trial version - purchase at www.fineprint.com

(α , ϕ)-Quantile Definition

- ϕ -skyline of α -skyline of P $\phi \cap$ P α



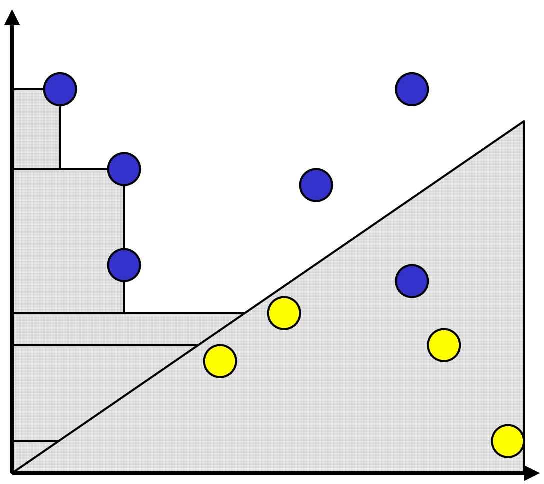
P α

Printed with FinePrint trial version - purchase at www.fineprint.com

(α , ϕ)-Quantile Definition

- ϕ -skyline of α -skyline of P $\phi \cap$ P α

P ϕ



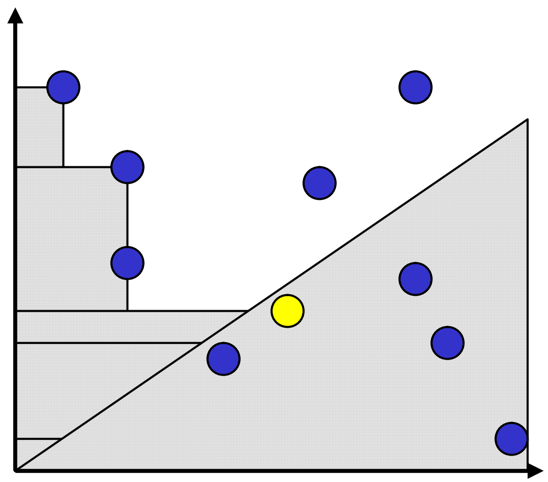
P α

Printed with FinePrint trial version - purchase at www.fineprint.com

(α , ϕ)-Quantile Definition

- ϕ -skyline of α -skyline of P $\phi \cap$ P α

P ϕ



P α

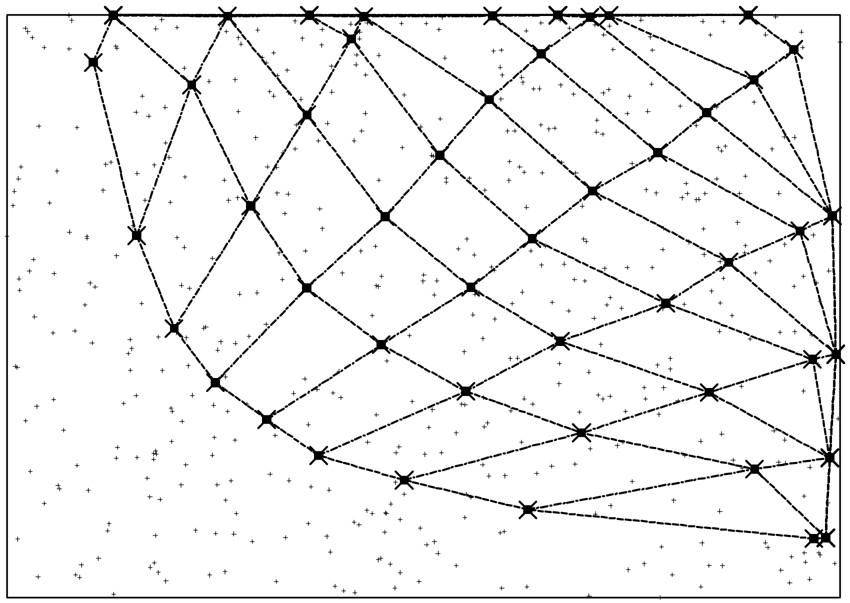
Printed with FinePrint trial version - purchase at www.fineprint.com

Properties of (α , ϕ)-Quantiles

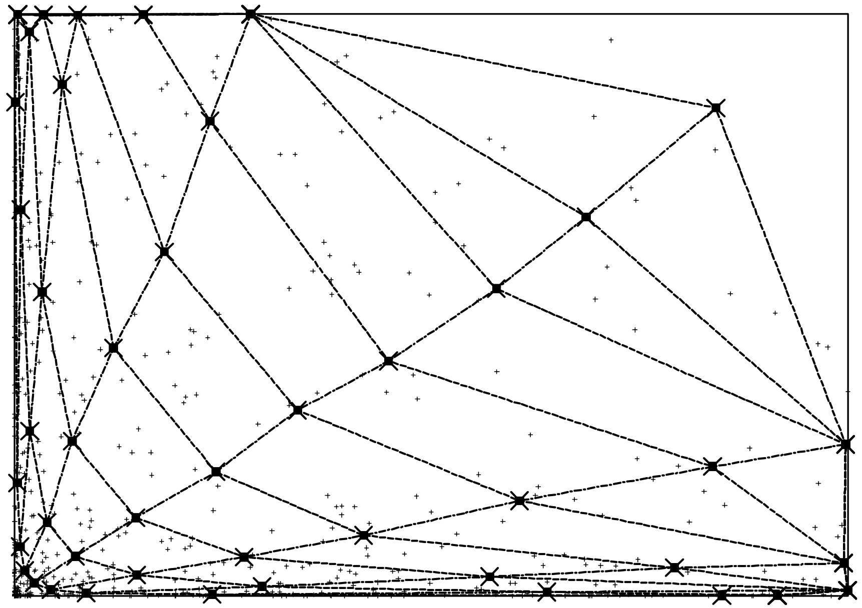
- Order of skylines on P ϕ ∩ P α doesn't matter
- p lies on α -radial or ϕ -quantour or both
- Uniqueness
  - Unique point p always found
  - p = q iff ϕ (p) = ϕ (q) and α (p) = α (q)
- Collapses to 1D quantiles for points (*x*,*x*)

Printed with FinePrint trial version - purchase at www.fineprint.com

Illustration



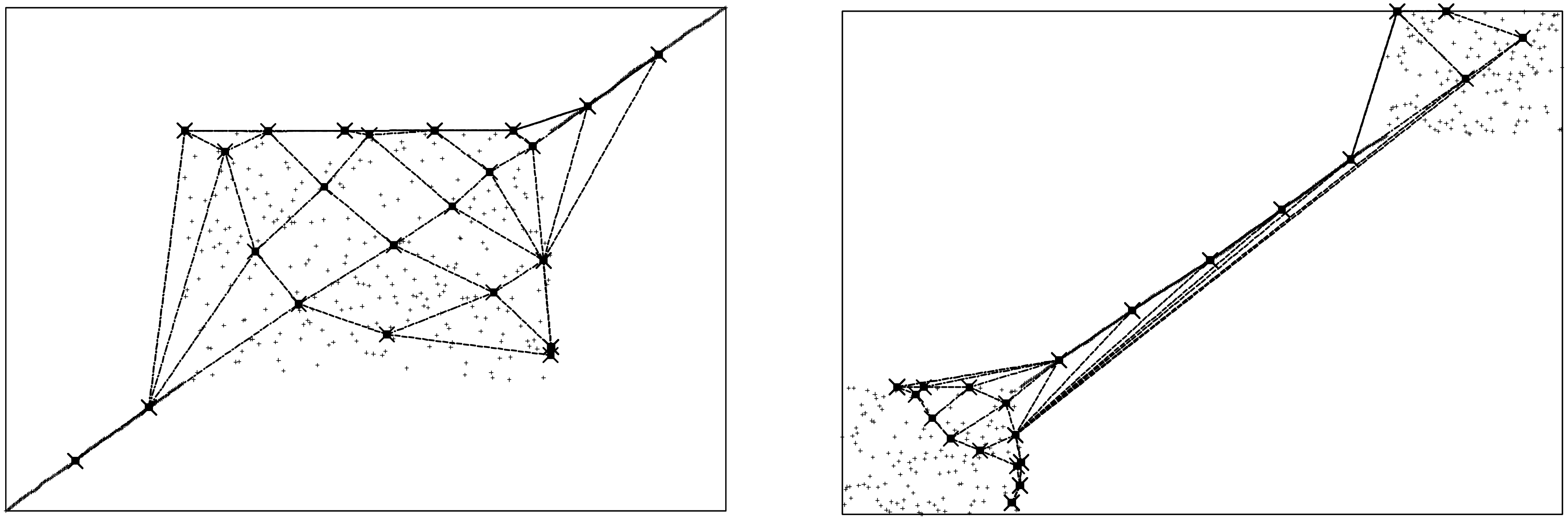
Uniform x Uniform



Zipf x Zipf

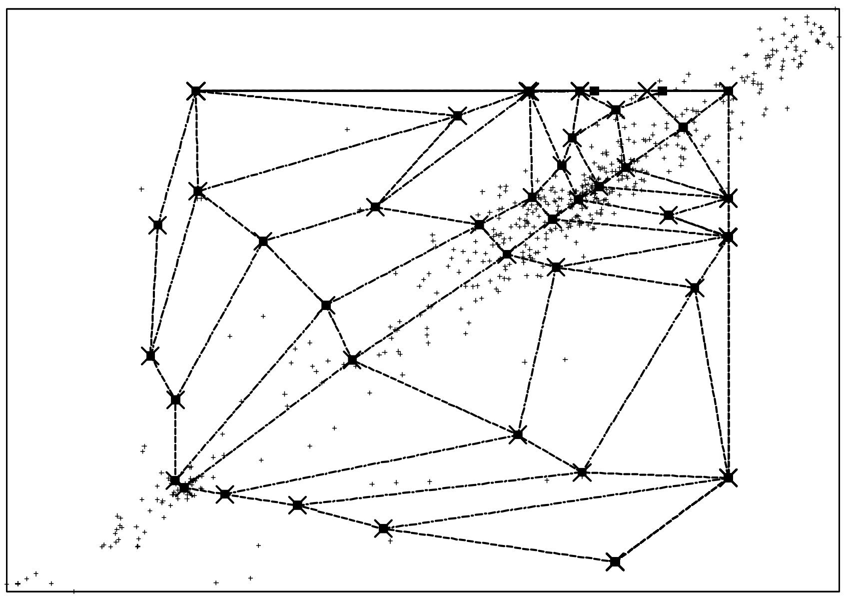
Printed with FinePrint trial version - purchase at www.fineprint.com

Illustration

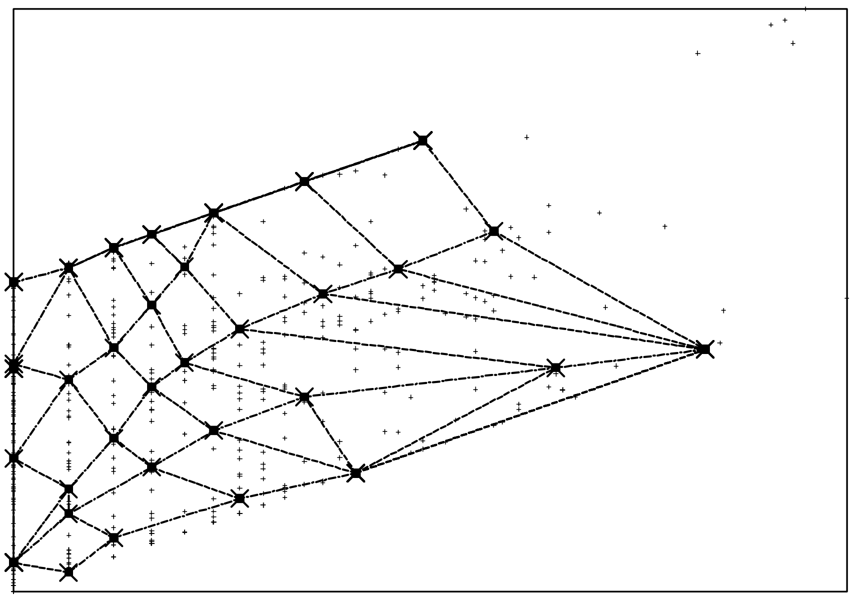


Printed with FinePrint trial version - purchase at www.fineprint.com

Illustration



SNMP data



Flow data (log-log)

Printed with FinePrint trial version - purchase at www.fineprint.com

Exact Algorithm

- Given (α , ϕ) find (i α ,j ϕ)-quantiles
- Pre-compute xrank(p), yrank(p), rank(p)
- Scan each point p and compute
  - i = Γ α (p) / α , j = Γ ϕ (p) / ϕ
  - Maintain point with max-*y* (max-*x* if tie) at (i,j)
  - Compute prefix max along rows/cols of (i,j)
- Complexity is O(N log N)

Printed with FinePrint trial version - purchase at www.fineprint.com

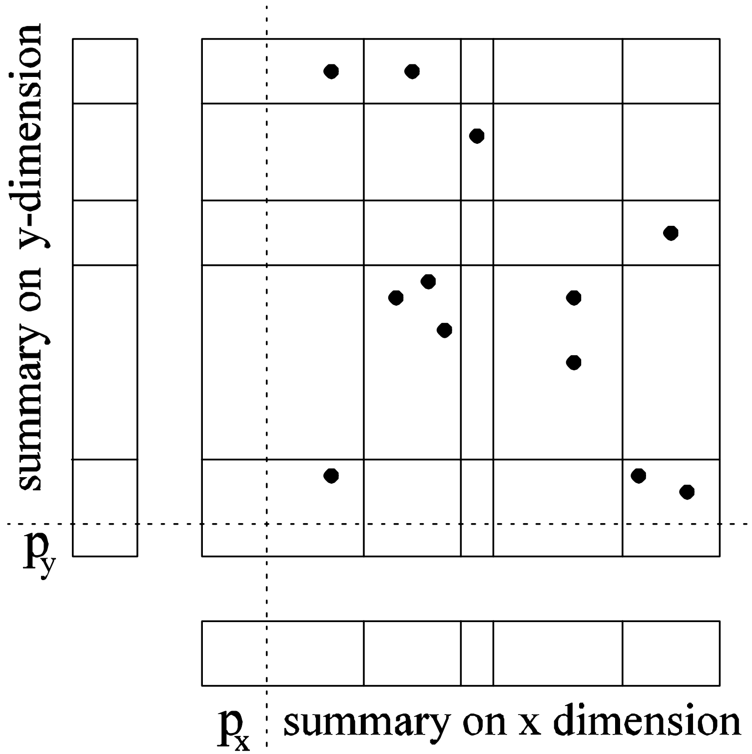
Streaming Algorithms

- Rank computation: given p, find (α (p), ϕ (p))
- Primitives: GK-digest, QD-digest, BQ-digest
- Three approaches
  - Cross-product
  - Deferred-merge
  - Eager-merge
- Two based on existing algorithms, one is novel
- Trade-offs (space, time, skew, etc)

Printed with FinePrint trial version - purchase at www.fineprint.com

Cross-Product Approach

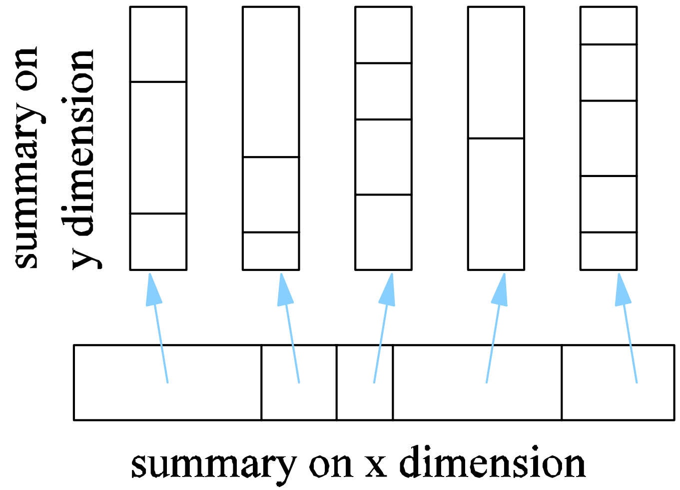
- 1D digest on each axis
- Cross-product grid
- Insert row + col
- Merge whole rows/cols
- GK x GK
  - – $O(1/\varepsilon^2 \log^2(\varepsilon N))$ space
  - – $O(1/\varepsilon \log \varepsilon N)$ update



Printed with FinePrint trial version - purchase at www.fineprint.com

Deferred-Merge Approach

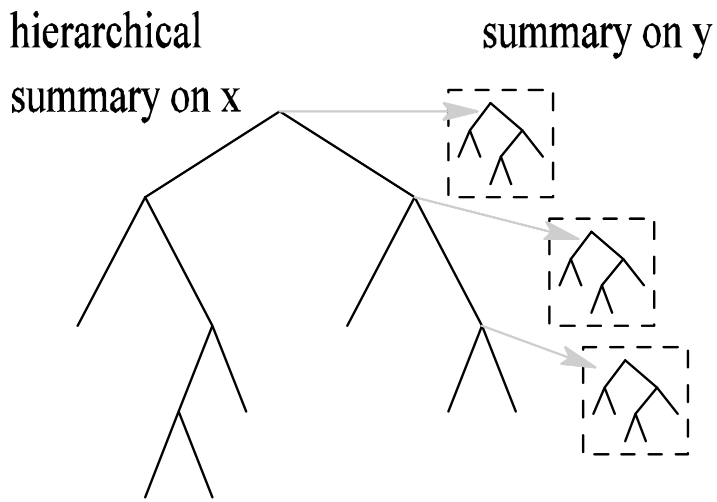
- Cascaded 1D digests
- QK x QD
  - $O(1/\varepsilon^2 \log(\varepsilon N)\log U)$ sp
  - $O(\log(1/\varepsilon \log U))$ update



Printed with FinePrint trial version - purchase at www.fineprint.com

Eager-Merge Approach

- Cascaded 1D Digests
- Multiple insertions
- No merging
- QD x QD
  - – $O(1/\varepsilon \log^3 U)$ space
  - – $O(\log U \log\log U)$ update



Printed with FinePrint trial version - purchase at www.fineprint.com

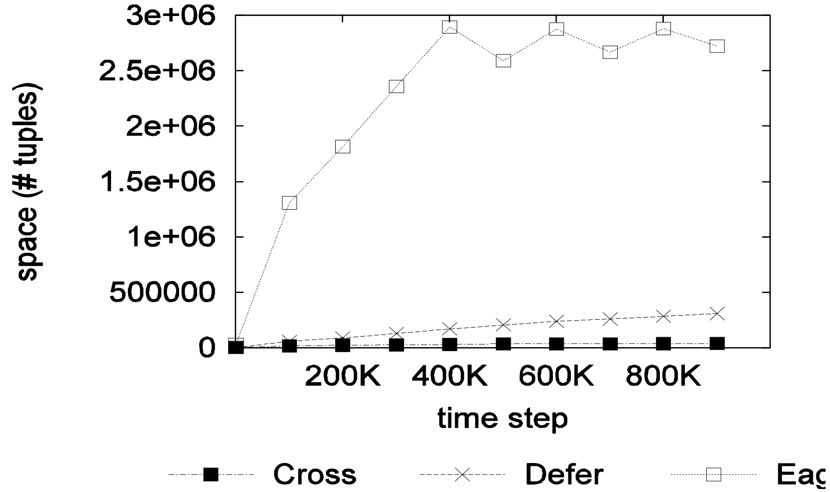
Summary of Results

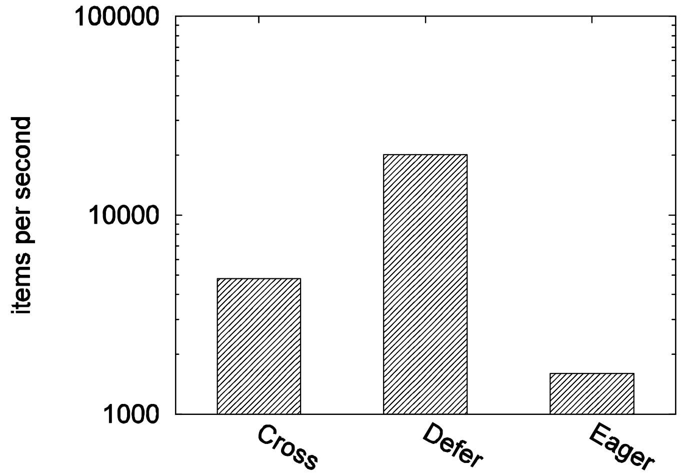
	Space	Update
Cross-Product (GK x GK)	$1/\varepsilon^2 \log^2(\varepsilon N)$	$1/\varepsilon \log(\varepsilon N)$
Deferred-Merge (GK x QD)	$1/\varepsilon^2 \log(\varepsilon N) \log U$	$\log(1/\varepsilon \log U)$
Eager-Merge (QD x QD)	$1/\varepsilon \log^3 U$	$\log U \log\log U$

Printed with FinePrint trial version - purchase at www.fineprint.com

Experiments

- Flow data
- ε = 0.01
- Space usage
  - Eager worst
- Throughput
  - Eager worst, defer best

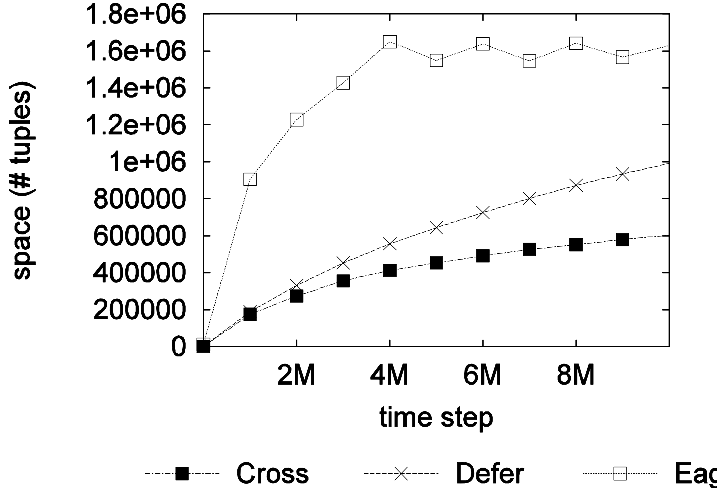


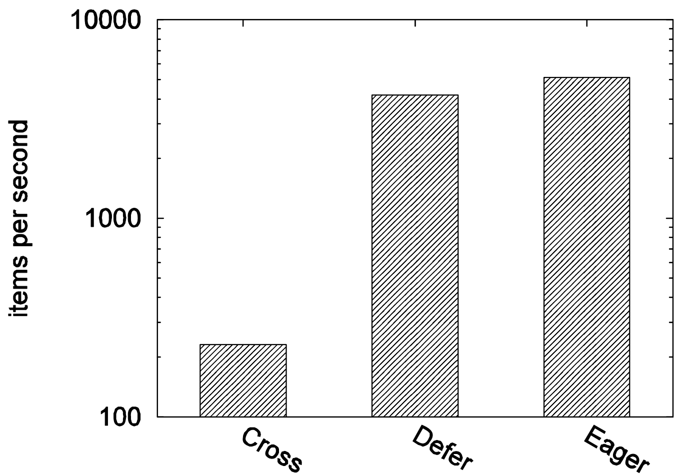


Printed with FinePrint trial version - purchase at www.fineprint.com

Experiments (cont'd)

- Flow data
- ε = 0.001
- Space usage
  - Eager worst
- Throughput
  - Cross-product worst





Printed with FinePrint trial version - purchase at www.fineprint.com

Conclusions

- Proposed (α , ϕ)-quantiles
  - Capture joint distribution (skew, dominance)
  - Natural definition, declarative semantics
  - Yield single points from the data
- Study of efficient streaming algorithms
  - Deferred-merge: novel and fast
- Future work
  - analysis for selection problem in streams

Printed with FinePrint trial version - purchase at www.fineprint.com

The End

Printed with FinePrint trial version - purchase at www.fineprint.com

Definitions

- p dominates q iff (px > qx) and (py > qy)
- P ϕ = points dominated by < ϕ N points
- ϕ -quantour: skyline of Pp
- skew = ranky(p) / [rankx(p) + ranky(p)]
- Pa = points with skew < aN points
- α -radial: skyline of Pa
- (a,p)-quantile: skyline of Pp ^ Pa

Printed with FinePrint trial version - purchase at www.fineprint.com