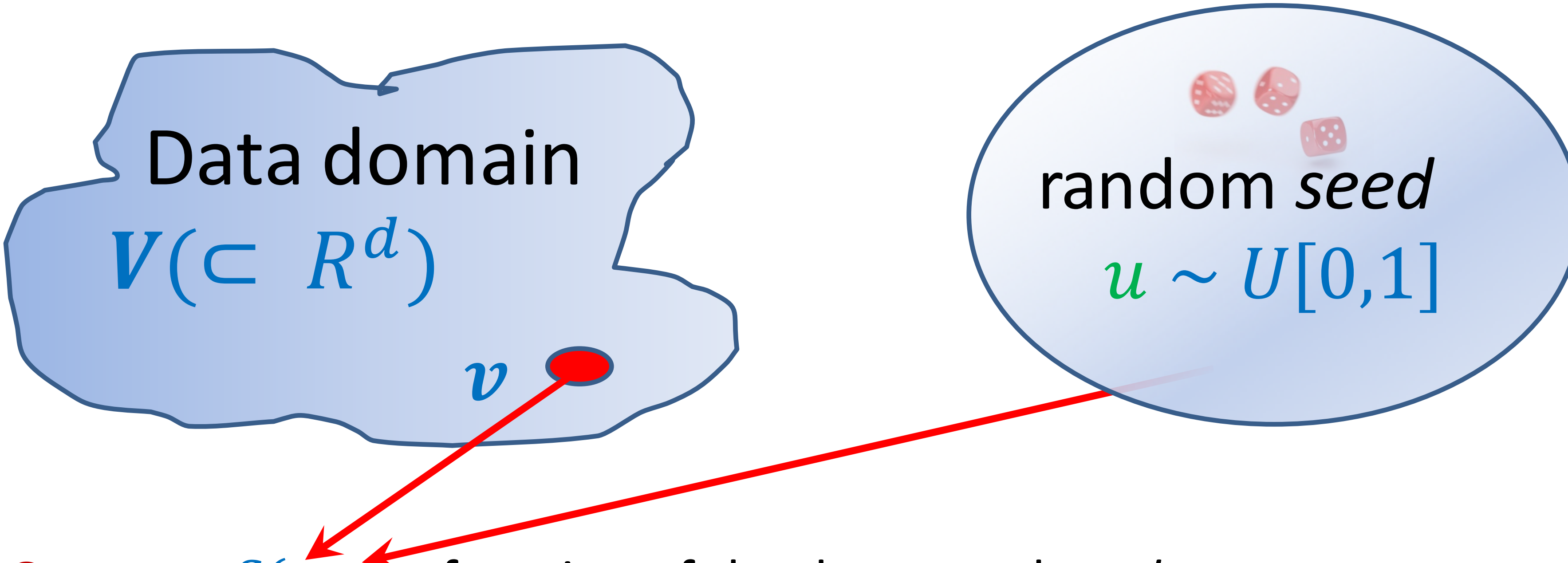
Monotone Estimation Framework and Applications for Scalable Analytics of Large Data Sets

Edith Cohen

Google Research

Tel Aviv University

A Monotone Sampling Scheme



Outcome $S(v, u)$: function of the data v and *seed* u

- *Seed value* u is available with the outcome

- $S(v, u)$ can be interpreted as the set of all data vectors consistent with the outcome and u

Monotonicity: Fixing v, $S(v, u)$ is non-increasing with u.

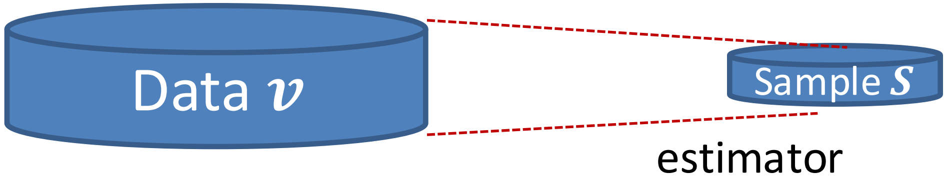
Monotone Estimation Problem (MEP)

A monotone sampling scheme $(\boldsymbol{V}, S)$:

- Data domain $\boldsymbol{V}(\subset \ R^d)$
- Sampling scheme $S: \boldsymbol{V} \times [0,1]$,

A nonnegative function $f: \boldsymbol{V} \ \geq 0$

Goal: **estimate $f(\boldsymbol{v})$**
from the sample



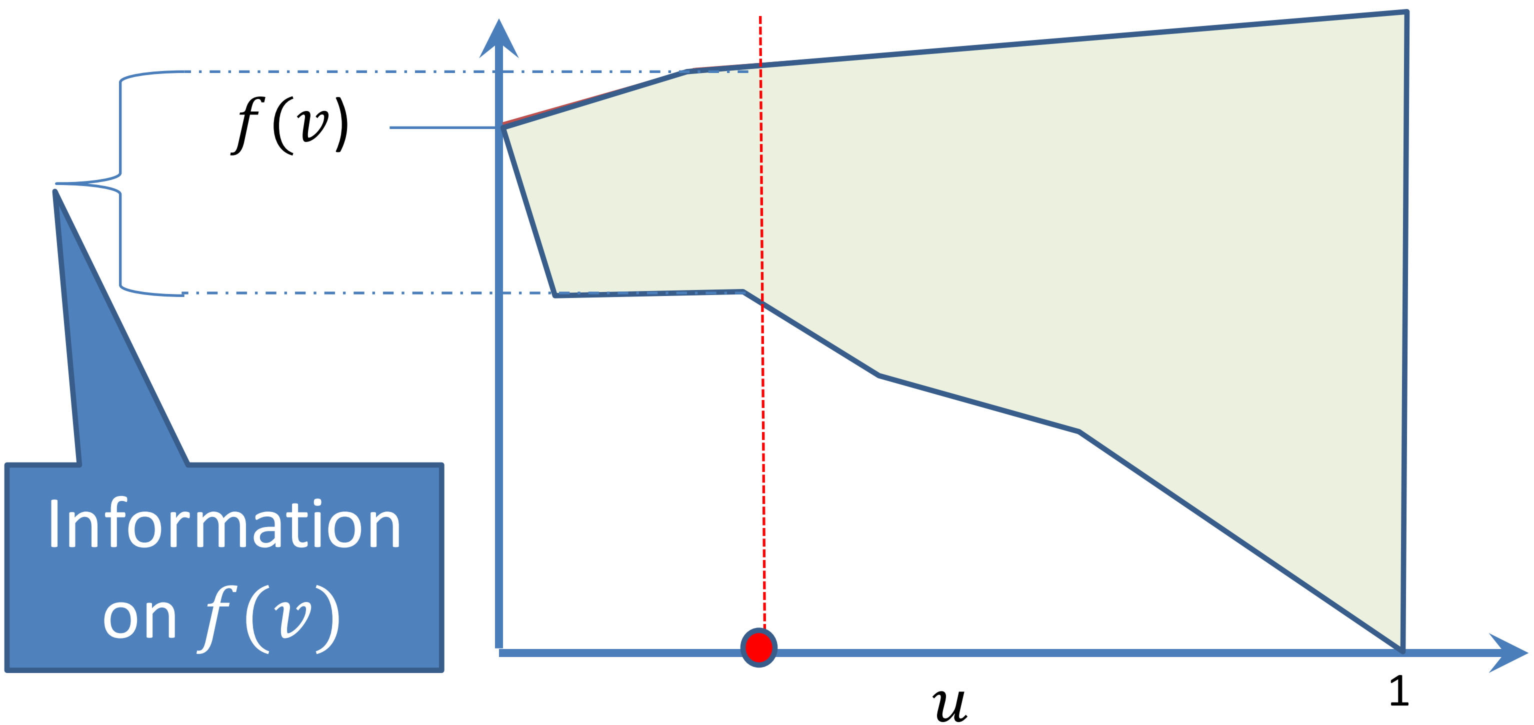
Q: $f(\boldsymbol{v})$ **?** $\longrightarrow$ $\hat{f}(\boldsymbol{S}(\boldsymbol{v},u),u)$

Desired properties of the estimator $\hat{f}(S)$:

- Unbiased $\forall \boldsymbol{v}, \int_u^1 \hat{f}(\boldsymbol{S}(\boldsymbol{v},x),x)dx = f(v)$ (useful with sums of MEPs)
- Nonnegative keep estimate $\hat{f}$ in the same domain as f
- (Pareto) "optimal" (admissible) any estimator with smaller $var_{u \sim U}[\hat{f}(S(\boldsymbol{v},u))]$ has for some $\boldsymbol{v}'$, larger $var_{u \sim U[0,]}[\hat{f}(S(\boldsymbol{v}',u))]$

Bounds on $f(\boldsymbol{v})$ from S and u

Data $\boldsymbol{v}$. The lower the seed u is, the more we know on $\boldsymbol{v}$ and hence on $f(\boldsymbol{v})$.



Estimators for MEPs

- **Unbiased, Nonnegative, Bounded variance**
- **Admissible:** "Pareto Optimal" in terms of variance

Results preview:

Explicit expressions for estimators for any MEP for which such estimator exists

Solution is not unique.

Consider some estimators with natural properties

- if we require monotonicity - $\hat{f}(S,u)$ is non-increasing with u, we get uniqueness

Notion of "Competitiveness" of estimators

We will come back to this, but first see some applications

MEP applications in data analysis

Scalable computation of approximate statistics and queries over large data sets

- Data is sampled (composable, distributed scheme). Sample is used to estimate statistics/queries expressed as a sum of multiple MEPs

- Key-value pairs with multiple sets of values (instances)
  - Take coordinated samples of instances. We get a MEP for each key
- Sketching graph-based influence and similarity functions
  - "Distance" sketch the utility values (relations of node to all others). Get a MEP for each "target" node from sketches of seed nodes
- Sketching generalized coverage functions
  - Coordinated weighted sample of the "utility" vector of each element. MEP for each item.

Social/Communication data

Activity value $v(x)$ is associated with each key $x = (b, c)$ (e.g. number of messages, communication from b to c)

- Take a weighted sample of keys. For example bottom-k ("weighed reservoir") or *PPS* (Probability Proportional to Size)

For $\boldsymbol{\tau} > 0$, iid $u(x) \sim U[0,1]$:

$$x \in S \leftrightarrow v(x) \geq \boldsymbol{\tau} \cdot u(x)$$

Monday activity	
(a,b)	40
(f,g)	5
(h,c)	20
(a,z)	10
……	
(h,f)	10
(f,s)	10

Monday Sample:	
(a,b)	40
(a,z)	10
……..	
(f,s)	10

- With bottom-k, $\boldsymbol{\tau}$ is set to obtain a fixed sample size k
- Without replacement sampling: $v(x) \geq -\boldsymbol{\tau} \ln u(x)$
- *Fully composable sampling scheme*

Samples of multiple days

Coordinated samples: Different values for different days.

Each key is sampled with *same seed $u(x)$ in different days*

Monday activity	
(a,b)	40
(f,g)	5
(h,c)	20
(a,z)	10
……	
(h,f)	10
(f,s)	10

Monday Sample:	
(a,b)	40
(a,z)	10
……..	
(f,s)	10

Tuesday activity	
(a,b)	3
(f,g)	5
(g,c)	10
(a,z)	50
……	
(s,f)	20
(g,h)	10

Tuesday Sample:
(g,c)
(a,z) 50
……..
(g,h)

Wednesday activity	
(a,b)	30
(g,c)	5
(h,c)	10
(a,z)	10
……	
(b,f)	20
(d,h)	10

Wednesday Sample:	
(a,b)	30
(b,f)	20
……..	
(d,h)	10

Matrix view keys $\times$ instances

In our example: keys $x = (a, b)$ are user pairs. Instances are days.

	Su	**Mo**	**Tu**	**We**	**Th**	**Fr**	**Sa**
(a,b)	40	30	10	43	55	30	20
(g,c)	0	5	0	0	4	0	10
(h,c)	5	0	0	60	3	0	2
(a,z)	20	10	5	24	15	7	4
(h,f)	0	7	6	3	8	5	20
(f,s)	0	0	0	20	100	70	50
(d,h)	13	10	8	0	0	5	6

Example Statistics

- Specify a segment of the keys $Y \subset X$, examples:
  - one user in CA and one in NY
  - apple device to android

Queries/Statistics $\sum_{x \in Y} f(v_1(x), v_2(x), \ldots, v_d(x))$

- Total communication of segment on Wednesday. $\sum_{x \in Y} v_1(x)$
- L_p^p distance/Weighted Jaccard change in activity of segment between Friday and Saturday $\sum_{x \in Y} |v_1(x) - v_2(x)|^p$
- L_p^p increase/decrease $\sum_{x \in Y} \max\{0, v_1(x) - v_2(x)\}^p$
- Coverage of segment Y in days D : $\sum_{x \in Y} \max_{i \in D} v_i(x)$
- Average/sum of median/max/min/top-3/concave aggregate of activity values over days D

We would like to compute an **estimate** from the **sample**

Matrix view keys $\times$ instances

Coordinated PPS sample $\tau = 100$ for all entries

u		Su	Mo	Tu	We	Th	Fr	Sa
0.33	(a,b)	40	30	10	43	55	30	20
0.22	(g,c)	0	5	0	0	4	0	10
0.82	(h,c)	5	0	0	60	3	0	2
0.16	(a,z)	20	10	5	24	15	7	4
0.92	(h,f)	0	7	6	3	8	5	20
0.16	(f,s)	0	0	0	20	100	70	50
0.77	(d,h)	13	10	8	0	0	5	6

Estimate sum statistics, one key at a time

$$\sum_{x \in Y} f(\boldsymbol{v}(\boldsymbol{x}))$$

Sum over keys $x \in Y$ of $f(\boldsymbol{v}(x))$, where $\boldsymbol{v}(x) = (v_1(x), v_2(x) \ldots)$

For L_p distance: $f(\boldsymbol{v}) = |v_1 - v_2|^p$

Estimate one key at a time:

$$\sum_{x \in Y} \hat{f}(S(v(x)))$$

The estimator for $f(\boldsymbol{v})$ is applied to the sample of $\boldsymbol{v}$

Easy statistics: Sum over entries Estimate a single entry at a time

- Example: Total communication of segment Y on Monday

Inverse probability estimate (Horviz Thompson) [HT52]:

Sum over sampled $x \in Y$ of $\frac{\boldsymbol{v_{monday}(x)}}{\boldsymbol{p_{monday}(x)}}$

Inclusion Probability $\boldsymbol{p_{monday}(x)}$ can be computed from $v(x)$ and $\boldsymbol{\tau}$:

$$x \in S \leftrightarrow v(x) \geq \boldsymbol{\tau} \cdot u(x)$$

$$p_i(x) = \Pr_{u \in U}[v_i(x) \geq \boldsymbol{\tau_i} \cdot u(x)]$$

HT estimator (single-instance)

Coordinated PPS sample $\tau = 100$

u
0.33
0.22
0.82
0.14
0.92
0.16
0.77

	Su	Mo	Tu	We	Th	Fr	Sa
(a,b)	40	30	10	43	55	30	20
(g,c)	0	5	0	0	4	0	10
(h,c)	5	0	0	60	3	0	2
(a,z)	20	10	5	24	13	7	4
(h,f)	0	7	6	3	8	5	20
(f,s)	0	0	0	20	100	70	50
(d,h)	13	10	8	0	0	5	6

HT estimator (single-instance)

$\tau = 100.$ Day: Wednesday, Segment: CA-NY

u
0.33
0.22
0.82
0.16
0.92
0.16
0.77

	Su	Mo	Tu	We	Th	Fr	Sa
(a,b)	40	30	10	43	55	30	20
(g,c)	0	5	0	0	4	0	10
(h,c)	5	0	0	60	3	0	2
(a,z)	20	10	5	24	15	7	4
(h,f)	0	7	6	3	8	5	20
(f,s)	0	0	0	20	100	70	50
(d,h)	13	10	8	0	0	5	6

HT estimator for single-instance

$\tau = 100.$ Day: Wednesday, Segment: CA-NY

u
0.33
0.22
0.82
0.16
0.92
0.16
0.77

	We
(a,b)	43
(g,c)	0
(h,c)	60
(a,z)	24
(h,f)	3
(f,s)	20
(d,h)	0

Exact: $43 + 60 + 20 = 123$

$p = 0.43$

HT estimate is 0 for keys that are not sampled, v/p when key is sampled

HT estimate: $100 + 100 = 200$

$p = 0.20$

Inverse-Probability (HT) estimator

- ✓ **Unbiased**: $(1 - p(x)) \cdot 0 + p(x)\frac{f(v(x))}{p(x)} = f(v(x))$
- ✓ **Nonnegative**: $v(x) \geq 0$ so $\frac{v(x)}{p(x)} \geq 0$
- ✓ **Bounded variance** (for all $\boldsymbol{v}$)
- ✓ **Monotone:** more information $\Rightarrow$ higher estimate
- ✓ **Optimal:** UMVU The unique minimum variance (unbiased, nonnegative, sum) estimator

Works when f depends on a single entry.
What about general f ?

Queries involving multiple columns

- L_p^p distance $\quad f(\boldsymbol{v}) = |v_1 - v_2|^p$
- L_p^p increase $\quad f(\boldsymbol{v}) = \max\{0, v_1 - v_2\}^p$

- HT estimate is positive only when we know $f(v) = |v_1 - v_2|$ from the sample.
- But for $v_2 = 0$, $v_1 > 0$ then $f(v) > 0$ but sample never reveals $f(v)$ because second entry is never sampled. Thus, HT is biased
- Even when unbiased, HT may not be optimal. E.g. when v_1 is sampled and we can deduce from τ_2 and u that $v_2 \leq a < v_1$ then we know that $f(v) \geq \mathrm{v}_1 - \mathrm{a}$. An optimal estimator will use this incomplete information
- We want estimators with the same nice properties as HT and optimality

Sampled data

Coordinated PPS sample $\tau = 100$

u
0.33
0.22
0.82
0.16
0.92
0.16
0.77

	Su	Mo	Tu	We	Th	Fr	Sa
(a,b)	40	30	10	43	55	30	20
(g,c)	0	5	0	0	4	0	10
(h,c)	5	0	0	60	3	0	2
(a,z)	20	10	5	24	15	7	4
(h,f)	0	7	6	3	8	5	20
(f,s)	0	0	0	20	100	70	50
(d,h)	13	10	8	0	0	5	6

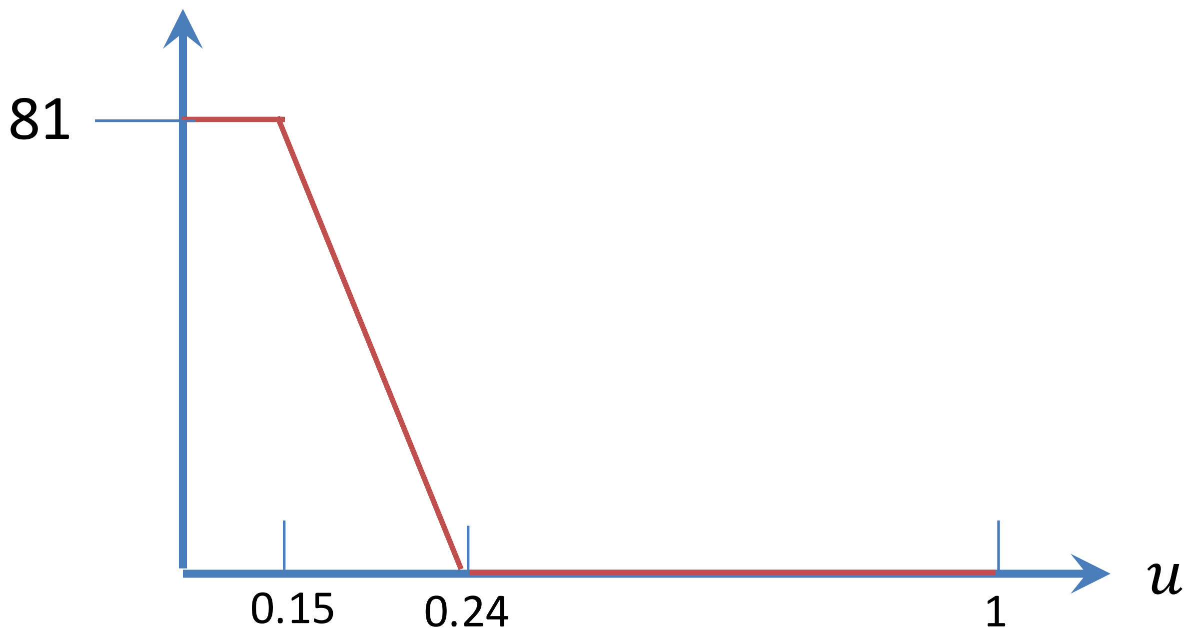
Want to estimate $(55-43)^2+(8-3)^2+(24-15)^2$

Lets look at key (a,z), and estimating $(24-15)^2$

Information on f

Fix the data $\boldsymbol{v}$. The lower u is, the more we know on $\boldsymbol{v}$ and on $f(\boldsymbol{v}) = (24 - 15)^2 = 81.$

We plot the lower bound we have on $f(\boldsymbol{v})$ as a function of the seed u.



This is a MEP !
Monotone Estimation Problem

A monotone sampling scheme $(\boldsymbol{V}, S)$:

- Data domain $\boldsymbol{V}(\subset\ R^d)$ here $(v_1, v_2) \in R^2_{\geq 0}$
- Sampling scheme $S: \boldsymbol{V} \times [0,1]$, here $S((v_1, v_2), u)$ reveals v_i when $v_i > 100\ u$

A nonnegative function $f: \boldsymbol{V}\ \geq 0$ here $(v_1 - v_2)^2$

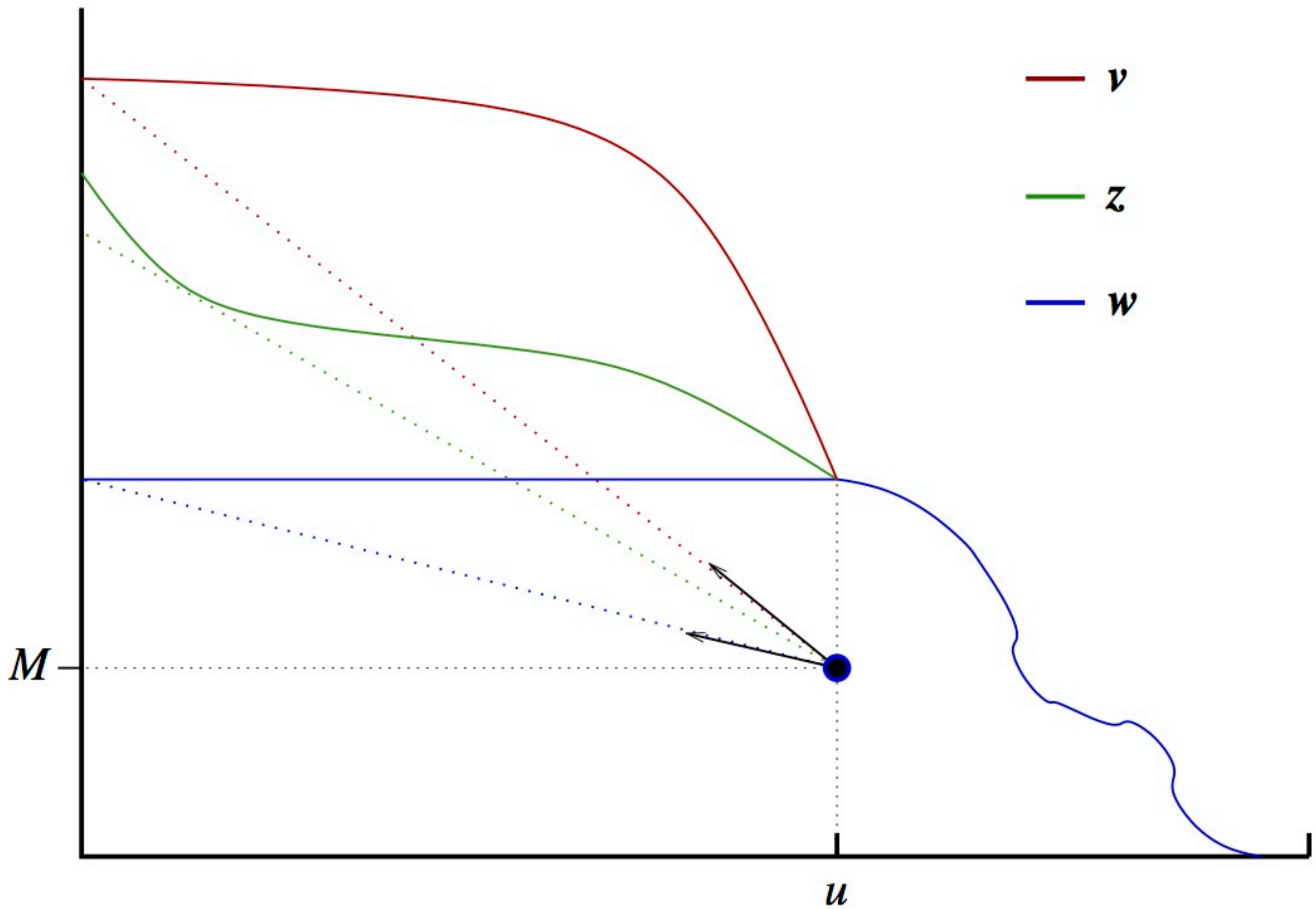
Goal: estimate $f(\boldsymbol{v})$: specify an *estimator* $\hat{f}(S, u)$ that is

Unbiased, Nonnegative, Bounded variance, Admissible (optimal)

Solution is not unique.

The optimal (admissible) range

Cum. Est
For $x > u$



We see $S(v,u)$ and u. We know what $S(v,x)$ is for all $x > u$.

Suppose we fixed $M = \int_u^1 \hat{f}(S(v,x),x)dx$

MEP Estimators

- **Order optimal estimators:** For an order $\prec$ on the data domain $\boldsymbol{V}$: Any estimator with lower variance on $\boldsymbol{v}$, must have higher variance on $\boldsymbol{z} \prec \boldsymbol{v}$

The L* estimator:

- The unique admissible **monotone** estimator
- Order optimal for: $\boldsymbol{z} \prec \boldsymbol{v} \Leftrightarrow \boldsymbol{f}(\boldsymbol{z}) < \boldsymbol{f}(\boldsymbol{v})$
- 4-variance competitive (soon we define that)

The U* estimator:

- Order optimal for: $\boldsymbol{z} \prec \boldsymbol{v} \Leftrightarrow \boldsymbol{f}(\boldsymbol{z}) > \boldsymbol{f}(\boldsymbol{v})$

Choice of estimator depends on properties we want, possibly depending on typical data distribution. L* is a good default (monotone and competitive)

Variance Competitiveness [CK13]

A "worst-case" over data theoretical indicator for estimator quality

For each $\boldsymbol{v}$, we can consider the minimum $E_{u \in U[0,1]}\,[\hat{f}^2(\mathrm{S}(\mathbf{v},\mathrm{u}), u)]$ attainable by an estimator that is unbiased and nonnegative for all other $\boldsymbol{v}$'

We use such "optimal" estimator $\hat{f}^{(v)}$ for $\boldsymbol{v}$ as a reference point.

An estimator $\hat{f}(S,u)$ is ***c-competitive*** if for any data $\boldsymbol{v}$, the expectation of the square is within a factor c of the minimum possible for $\boldsymbol{v}$ (by an unbiased and nonnegative estimator).

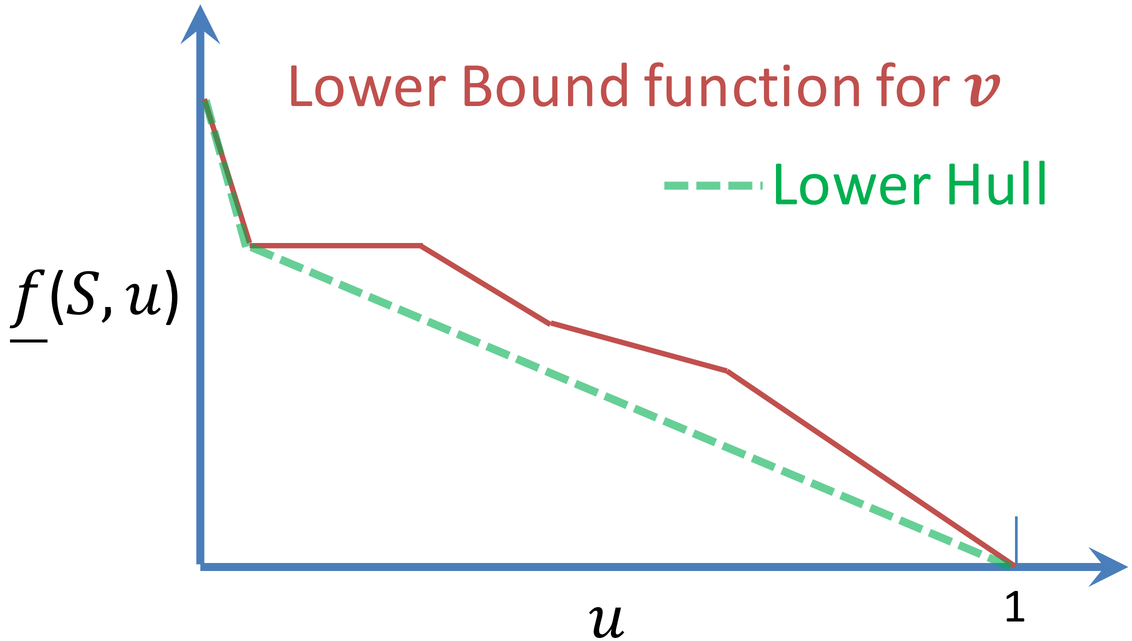
For all unbiased nonnegative $\hat{g}$,

$$E_{u \in U[0,1]}\,[\hat{f}^2(\mathrm{S}(\mathbf{v},\mathrm{u}))] \le c\;\; E_{u \in U[0,1]}\,[\hat{g}^2(\mathrm{S}(\mathbf{v},\mathrm{u}))]$$

The L^* estimator is 4-competitive and this is tight. For some MEPs, ratio is 4

Optimal estimator $\hat{f}^{(\boldsymbol{v})}$ for data $\boldsymbol{v}$ (unbiased and nonnegative for all data)

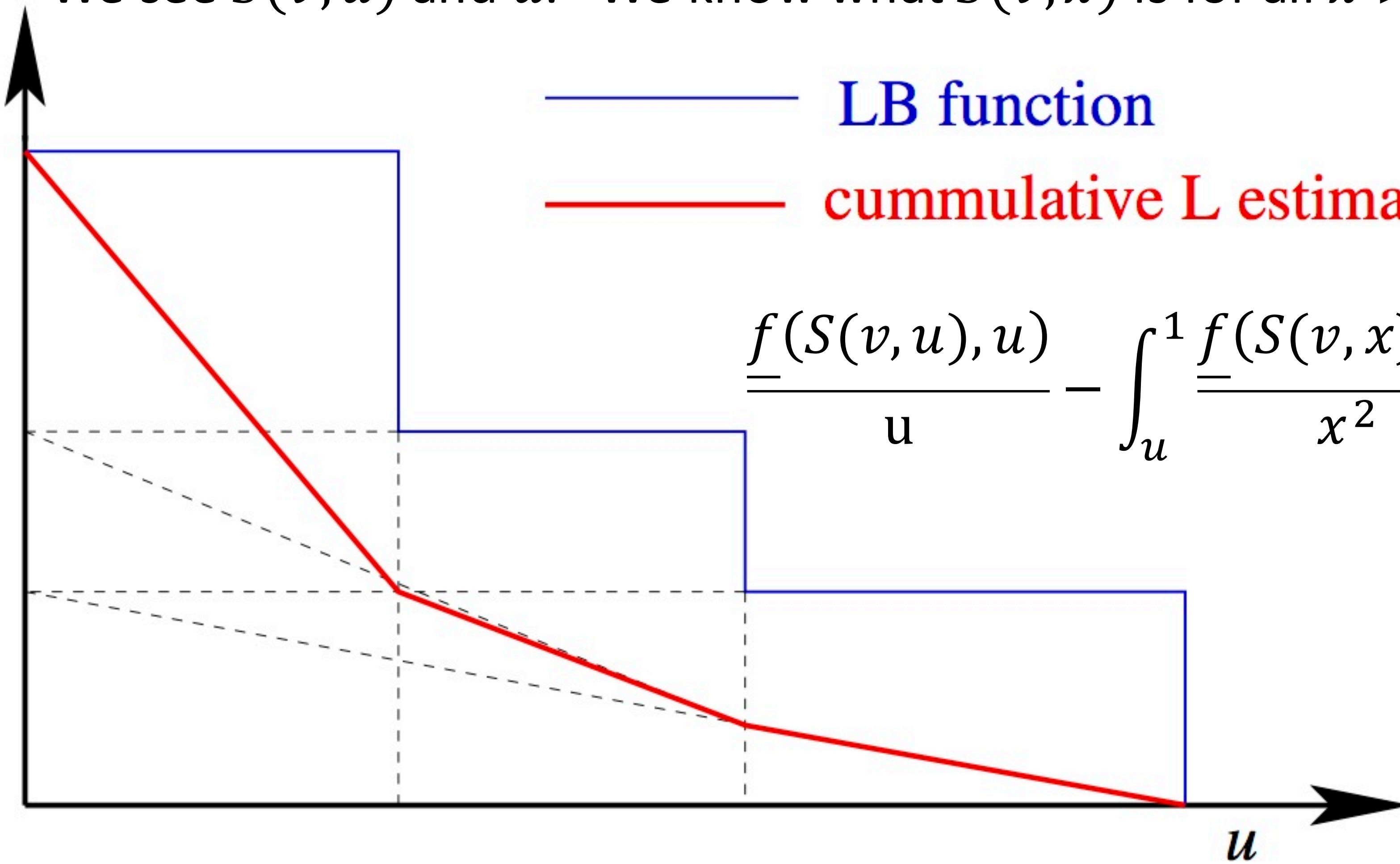
The optimal estimates $\hat{f}^{(\boldsymbol{v})}$ are the negated derivative of the lower hull of the Lower bound function.



Intuition: The lower bound guides us on outcome S, how "high" we can go with the estimate, in order to optimize variance for $\boldsymbol{v}$ while still being nonnegative on all other consistent data vectors.

The L* estimator

We see $S(v,u)$ and u. We know what $S(v,x)$ is for all $x > u$



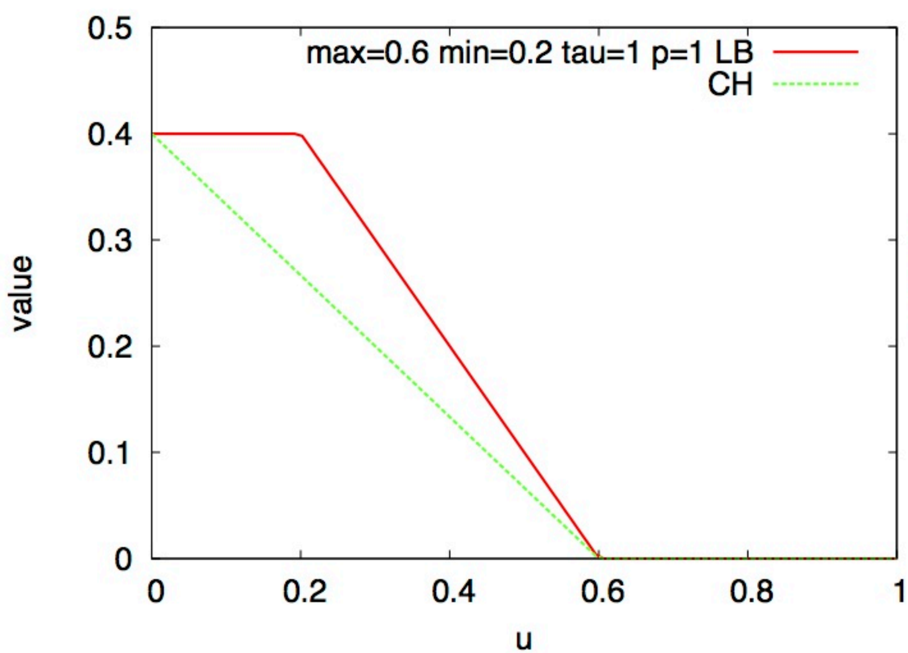
$$\frac{\underline{f}(S(v,u),u)}{u} - \int_{u}^{1} \frac{\underline{f}(S(v,x),x)}{x^2}\,dx$$

L_1 estimation example

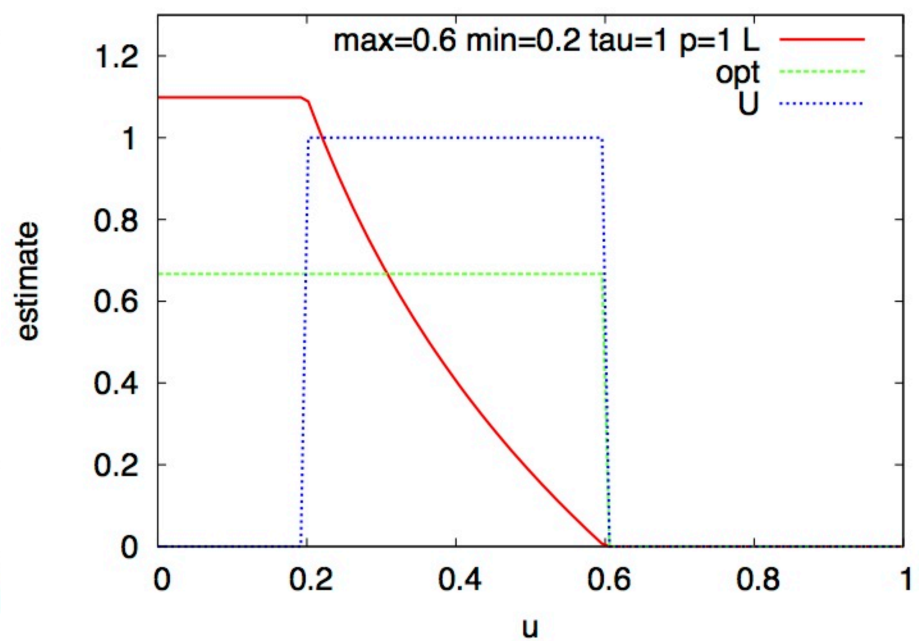
Estimators for $f(\mathrm{v}_1, \mathrm{v}_2) = |v_1 - v_2|$

Scheme: $v_i \geq 0$ is sampled if $v_i > u$

- "lower bound" (LB) on $f(0.6{,}0.2\ )$ from S and u
- The Lower hull of LB



The L^*, U^*, and opt for v estimators



U^* is optimized for the vector $f(0.6{,}0.0)$ (always consistent with S)

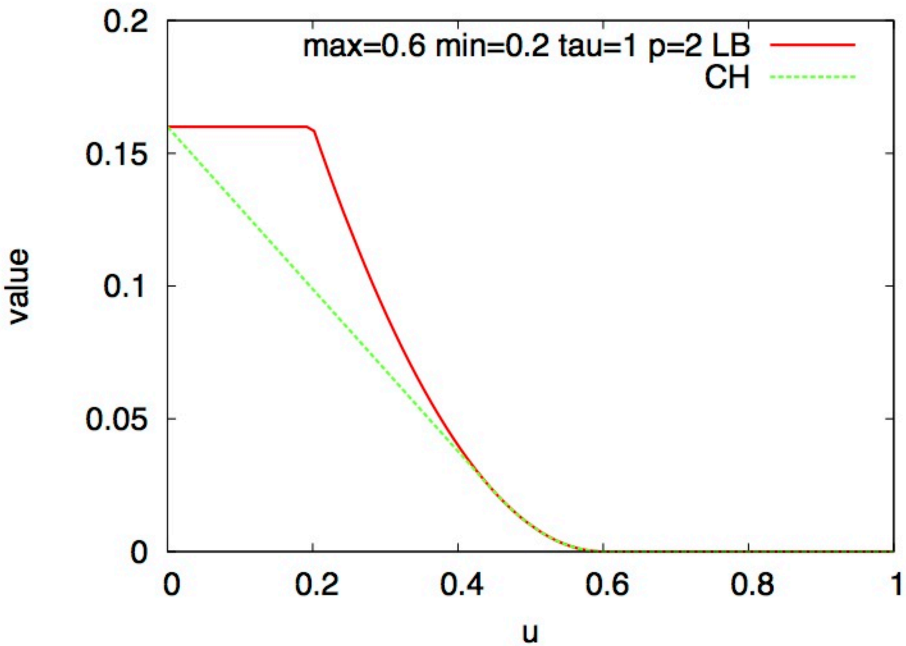
L^* is optimized locally for the vector $f(0.6, u)$ (consistent vector with smallest f

L_2^2 estimation example

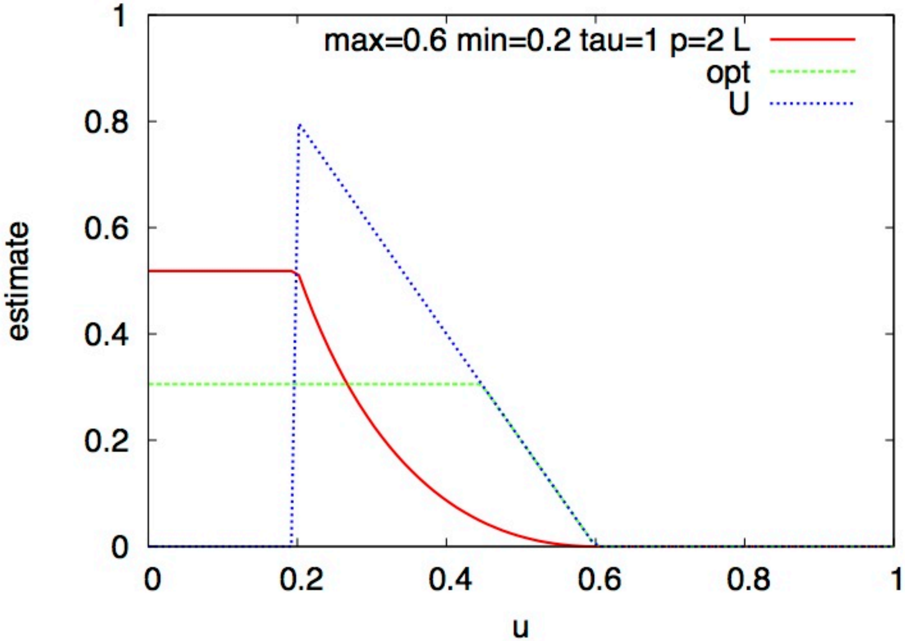
Estimators for $f(\mathrm{v}_1, \mathrm{v}_2) = |v_1 - v_2|^2$

Scheme: $v_i \geq 0$ is sampled if $v_i > u$

- "lower bound" (LB) on $f(0.6, 0.2)$ from S and u
- The Lower hull of LB



The L^*, U^*, and opt for v estimators



U^* is optimized for the vector $f(0.6, 0.0)$ (always consistent with S)

L^* is optimized locally for the vector $f(0.6, u)$ (consistent vector with smallest f

Summary

- Defined Monotone Estimation Problems (MEPs) (motivated by coordinated sampling)
- Derive Pareto optimal (admissible) unbiased and nonnegative estimators (for any MEP when they exist):
  - L* (lower end of range: unique monotone estimator, dominates HT) ,
  - U* (upper end of range),
  - Order optimal estimators (optimized for certain data patterns)

Applications

- Estimators for Euclidean and Manhattan distances from samples [C KDD '14]
- sketch-based closeness similarity in social networks [CDFGGW COSN '13] (similarity of the sets of long-range interactions)
- Sketching generalized coverage functions, including graph-based influence functions [CDPW '14, C' 16]

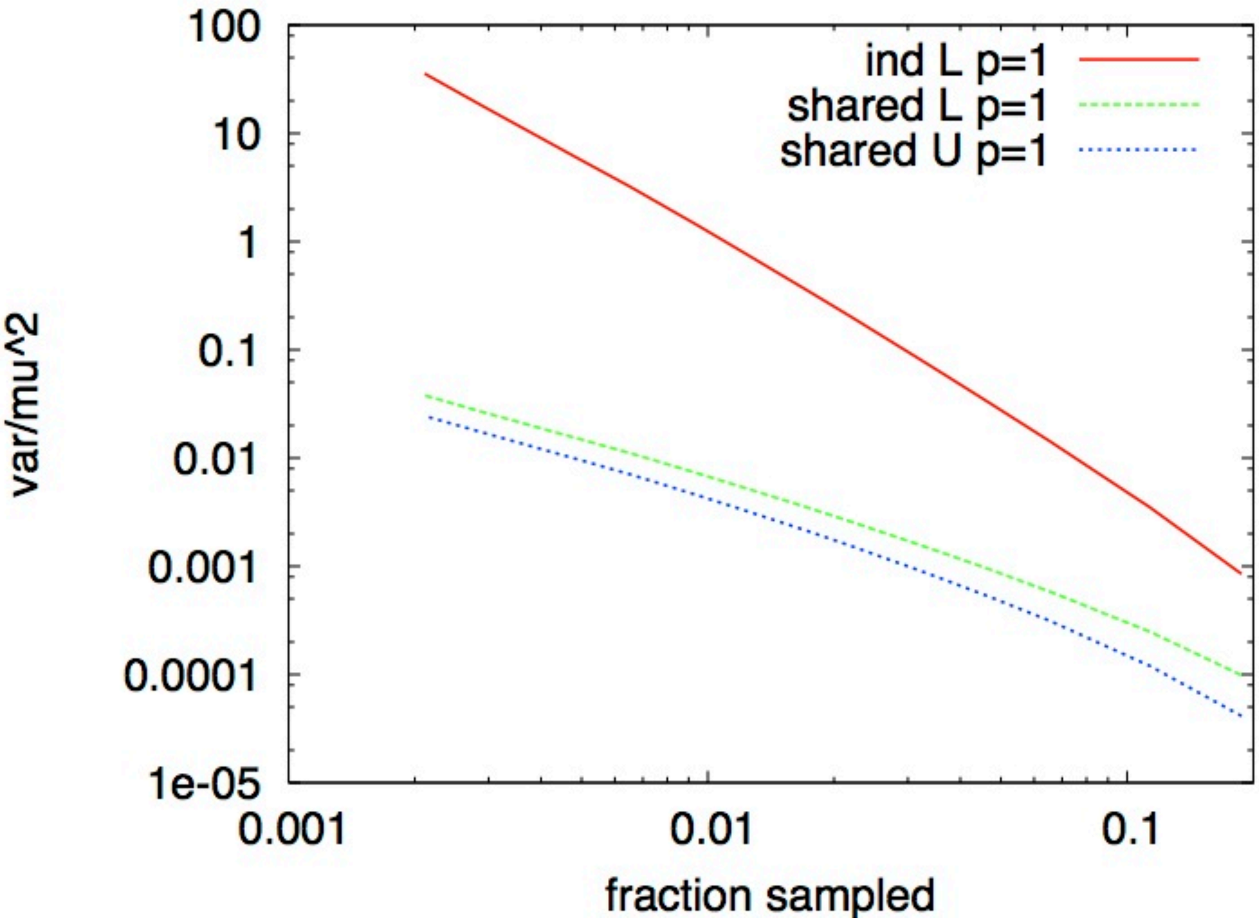
Future

- Tighter bounds on universal ratio: L* is 4 competitive, can do 3.375 competitive, lower bound is 1.44 competitive.
- Instance-optimal competitiveness – **Give efficient construction for any MEP**
- Multi-dimensional MEPs: Have multiple independent seed (independent samples of "columns"), some initial derivations for $d = 2$ and coverage and distance functions [CK 12, C 14], but the full picture is missing

L_1 distance [C KDD14]

Independent / Coordinated PPS sampling

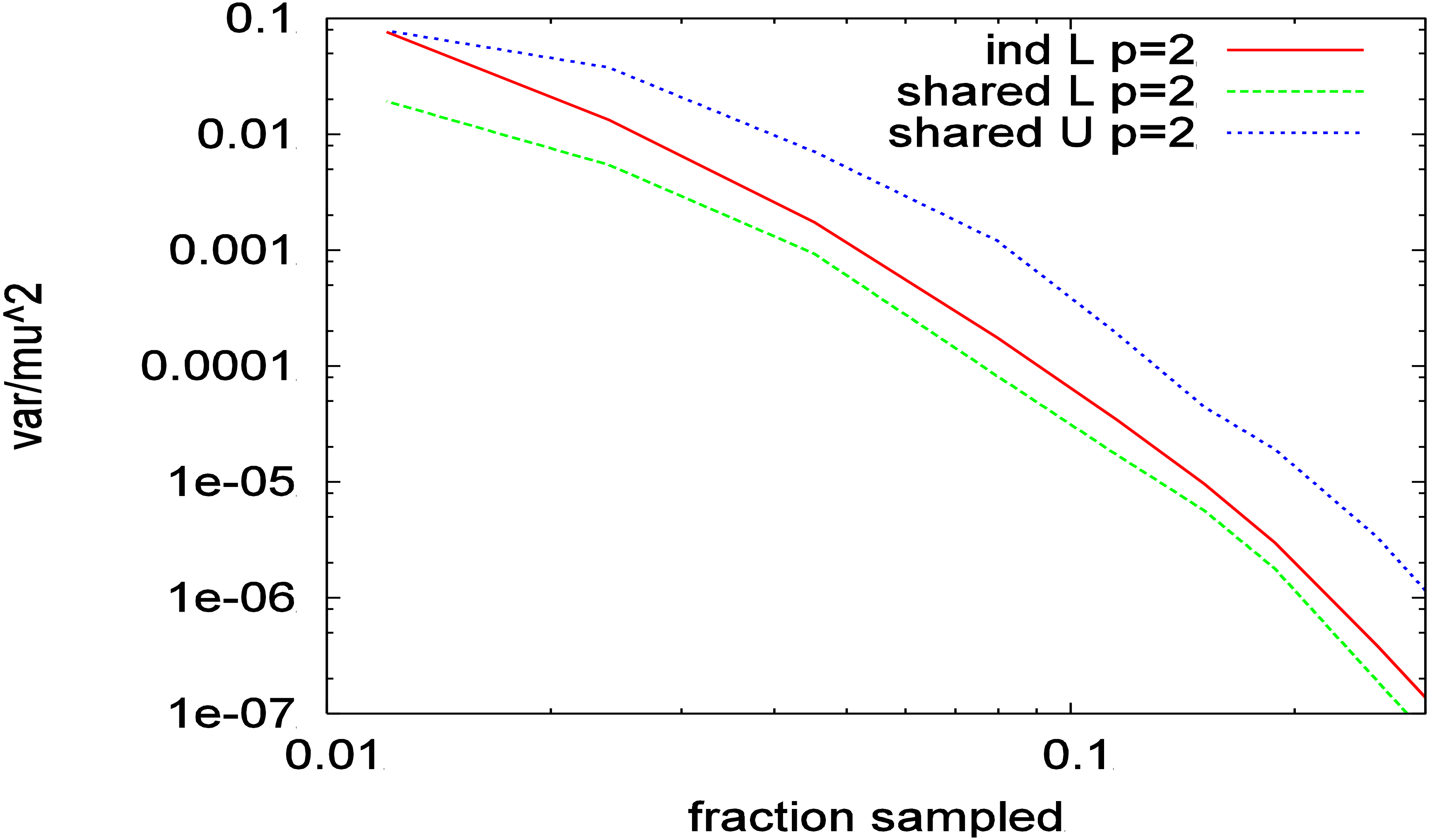
#IP flows to a destination in two time periods



L_2^2 distance [C KDD14]

Independent/Coordinated PPS sampling

Surname occurrences in 2007, 2008 books (Google ngrams)



Thank you!

Coordination of samples

Very powerful tool for big data analysis with applications well beyond what [Brewer, Early, Joyce 1972] could envision

- **Locality Sensitive Hashing** (LSH) (similar weight vectors have similar samples/sketches)
- **Multi-objective samples** (universal samples): A single sample (as small as possible) that provides statistical guarantees for multiple sets of weights.
- **Statistics/Domain queries that span multiple "instances"** (Jaccard similarity, L_p distances, distinct counts, union size,…)
  - MinHash sketches are a special case with 0/1 weights.
- **Facilitates faster computation of samples**. Example: [C' 97] Sketching/sampling reachability sets and neighborhoods of all nodes in a graph in near-linear time.