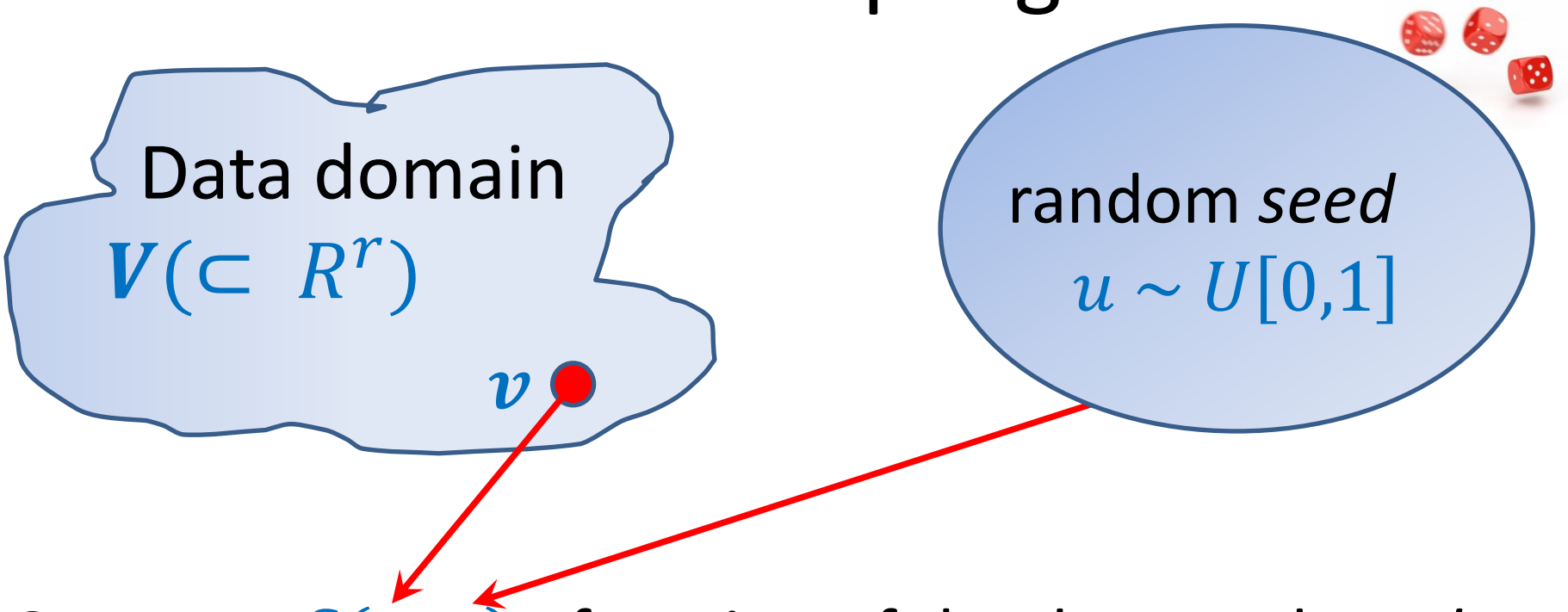
Estimation for Monotone Sampling: Competitiveness and Customization

Edith Cohen

Microsoft Research

Microsoft Research

A Monotone Sampling Scheme



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

Monotone: Fixing $\boldsymbol{v}$ the *information* in $S(u,v)$ (set of all data vectors consistent with S and u) is non-increasing with u.

Monotone Estimation Problem (MEP)

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

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

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

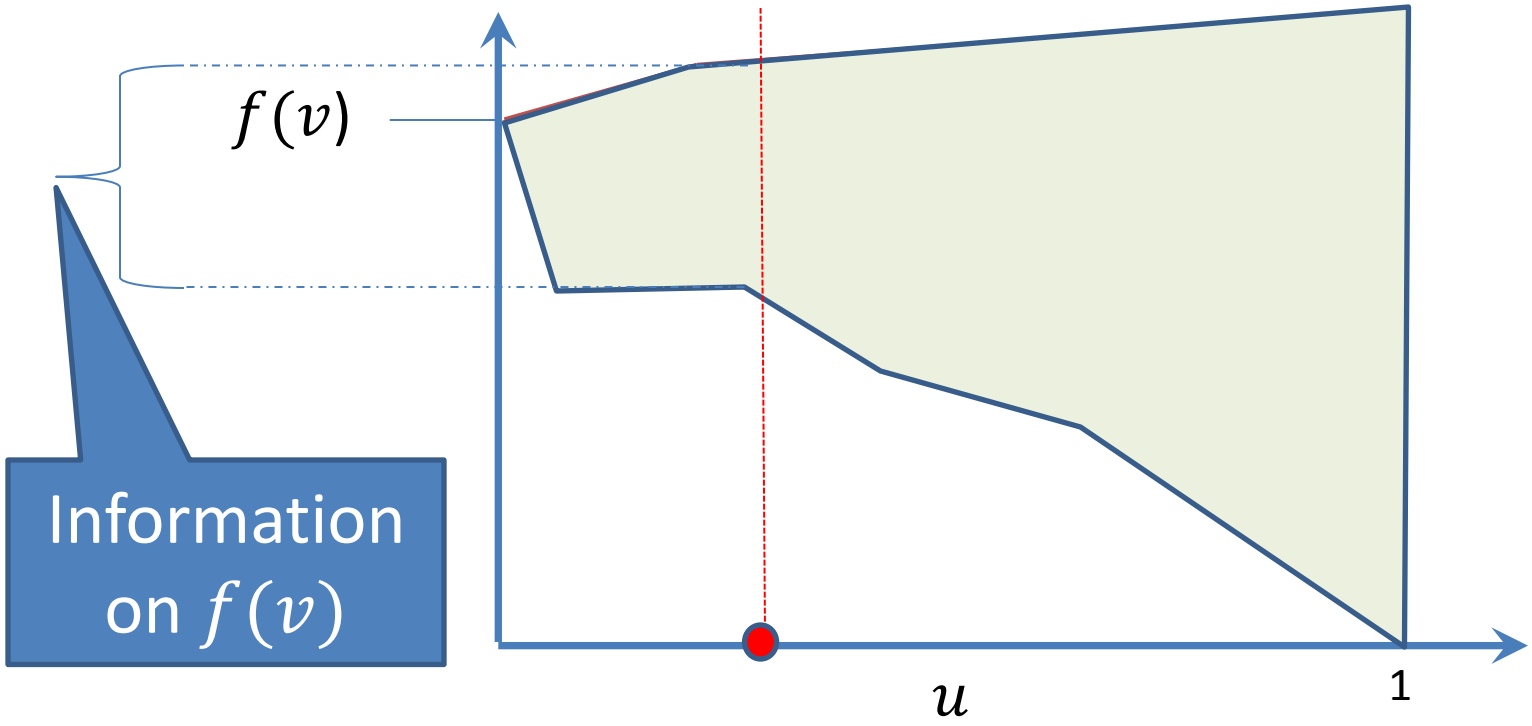
Goal: **estimate** $f(\boldsymbol{v})$

Specify an *estimator* $\hat{f}(S)$ that is:

Unbiased, nonnegative, (Pareto) "optimal"

What we know on $f(\boldsymbol{v})$ from S

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



MEP applications in data analysis:

Scalable but perhaps approximate query processing:

Data is sampled/sketched/summarized.
We process queries posed over the data by applying an estimator to the sample.

We give an example

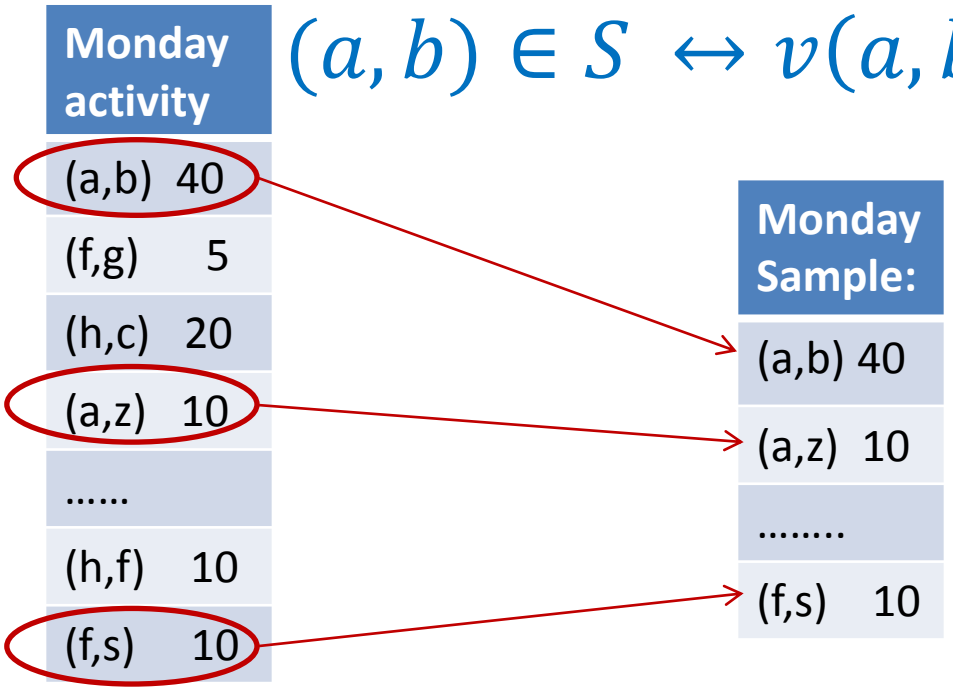
Example: Social/Communication data

Activity value $v(b,c)$ is associated with each node pair (b,c) (e.g. number of messages, communication)

Pairs are *PPS sampled* (Probability Proportional to Size)

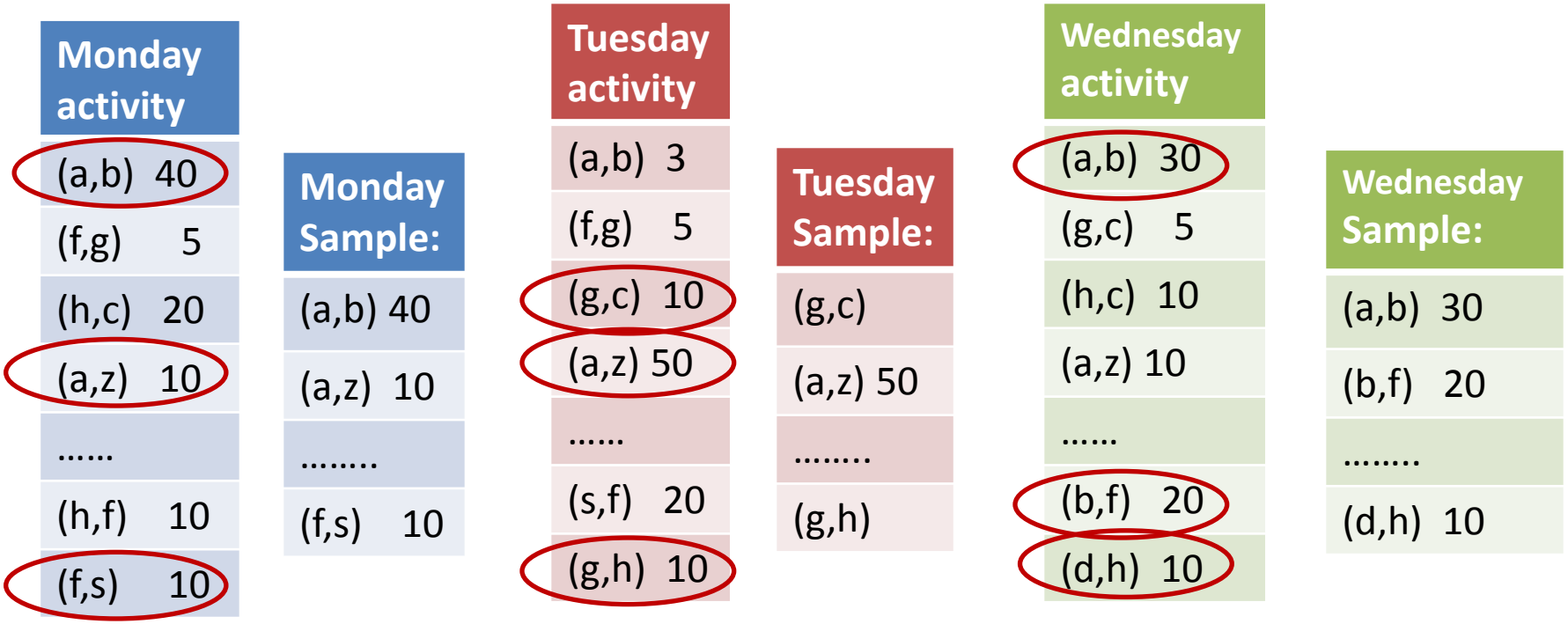
For $\tau > 0$, iid $u(a,b) \sim U[0,1]$:

$$(a,b) \in S \leftrightarrow v(a,b) \geq \tau \cdot u(a,b)$$



Samples of multiple days

Coordinated samples: Each pair is sampled with *same seed* $u(a,b)$ *in different days*



Matrix view keys × instances

In our example: keys (a,b) are user-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

Matrix view keys × instances

Coordinated PPS sample $\tau = 100$

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

Example Queries

- Total communication from users in California to users in New York on Wednesday.
- L_p distance (change) in activity of male-male users over 30 between Friday and Monday
- Breakdown: total increase, total decrease
- Average of median/max/min activity over days

We would like to **estimate** the query result from the **sample**

Estimate one key at a time

Queries are often (functions of) **sums** over **selected keys** h of a function f applied to the values tuple of h

$$\boldsymbol{v}^{(h)} = (v^{(h)}{}_1, v^{(h)}{}_2, v^{(h)}{}_3, \dots)$$

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

Estimate one key at a time:

$$\sum_h \hat{f}(S_h)$$ ⟸ The estimator for $f(\boldsymbol{v}^{(h)})$ is applied to the sample of $\boldsymbol{v}^{(h)}$

“Warmup” queries: Estimate a single entry at a time

- Total communication from users in California to users in New York on Wednesday.

Inverse probability estimate

(Horviz Thompson) [HT52]:

Over sampled entries $\boldsymbol{h}$ that match predicate (CA to NY, Wednesday), add up value divided by inclusion probability in sample $\boldsymbol{v(h)/p(h)}$

HT estimator (single-instance)

Coordinated PPS sample $\tau = 100$

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
0.14	(a,z)	20	10	5	24	13	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

HT estimator (single-instance)

$\tau = 100$. Select Wednesday, CA-NY

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

HT estimator for single-instance

$\tau = 100$. Select Wednesday, 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**: important because bias adds up and we are estimating sums
- **Nonnegative**: important because f is
- **Bounded variance** (for all $\boldsymbol{v}$)
- **Monotone:** more information ⇒ higher estimate
  - **Optimality:** 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 distance (change) in activity of "male users over 30" between Friday and Monday

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

- Breakdown: total increase, total decrease

$$f(\boldsymbol{v}) = \max\{0, v_1 - v_2\}^p$$

HT may not work at all now and may not be optimal when it does.

We want estimators with the same nice properties

Sampled data

Coordinated PPS sample $\tau = 100$

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

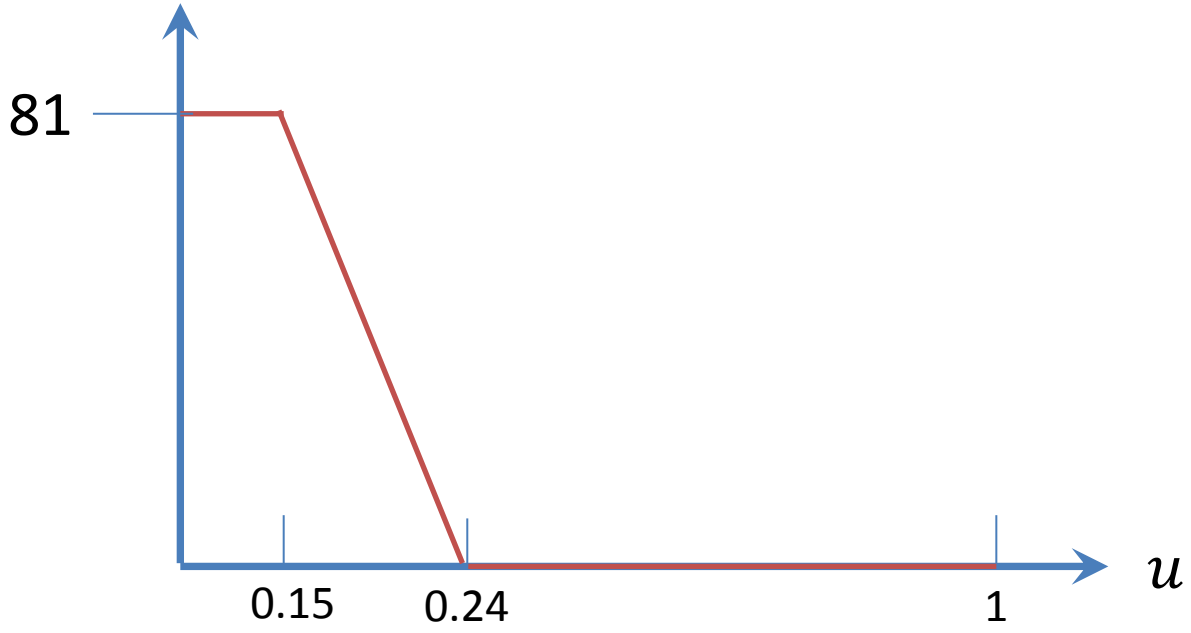
Want to estimate $(55-43)^2 + (8-3)^2 + (24-15)^2$

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

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

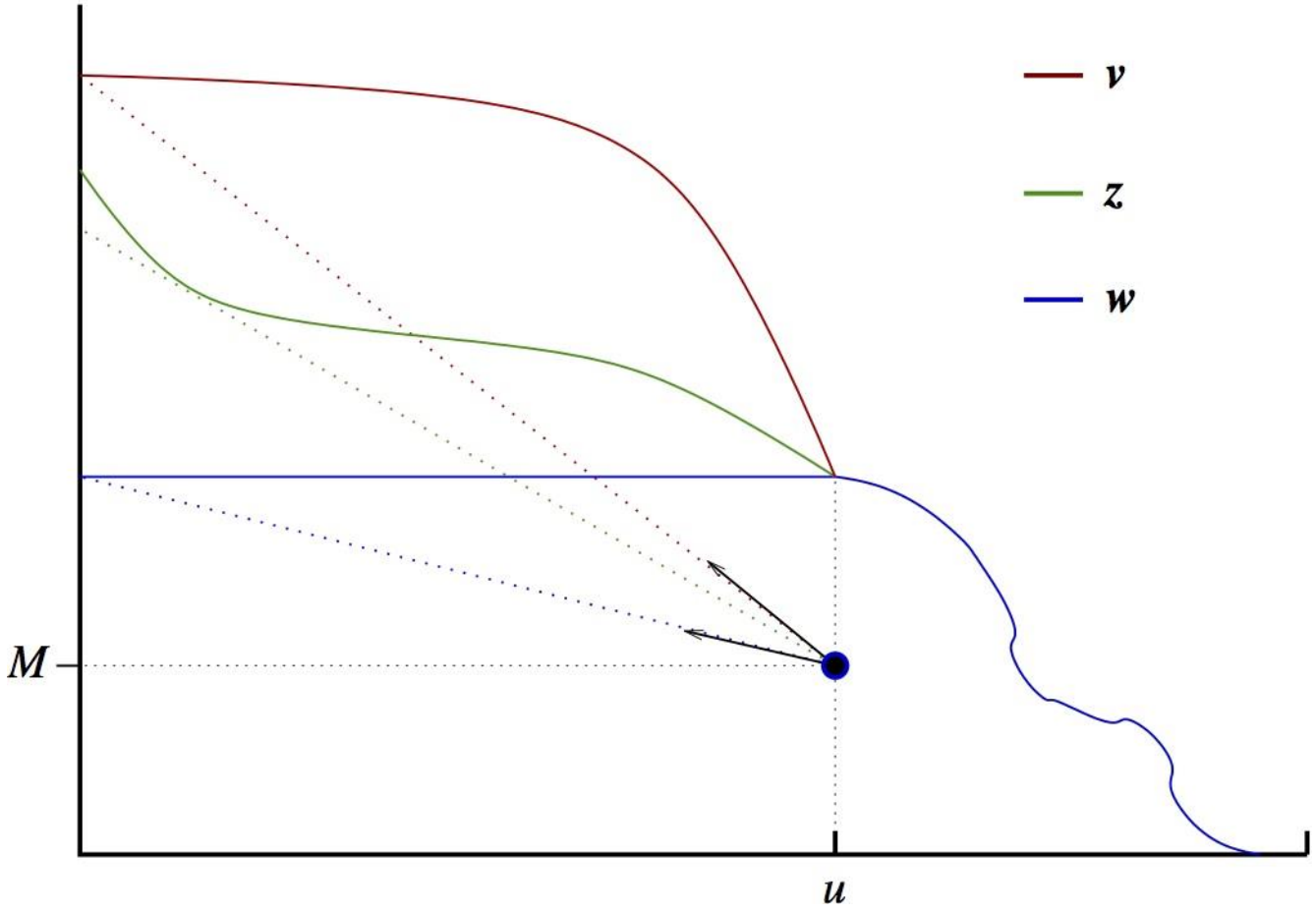
Goal: estimate $f(\boldsymbol{v})$: specify a good *estimator* $\hat{f}(S, u)$

Our results:
General Estimator Derivations for any MEP for which such estimator exists

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

Solution is not unique.

The optimal range



Our results:
General Estimator Derivations

- **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 $\boldsymbol{z} \prec \boldsymbol{v}$

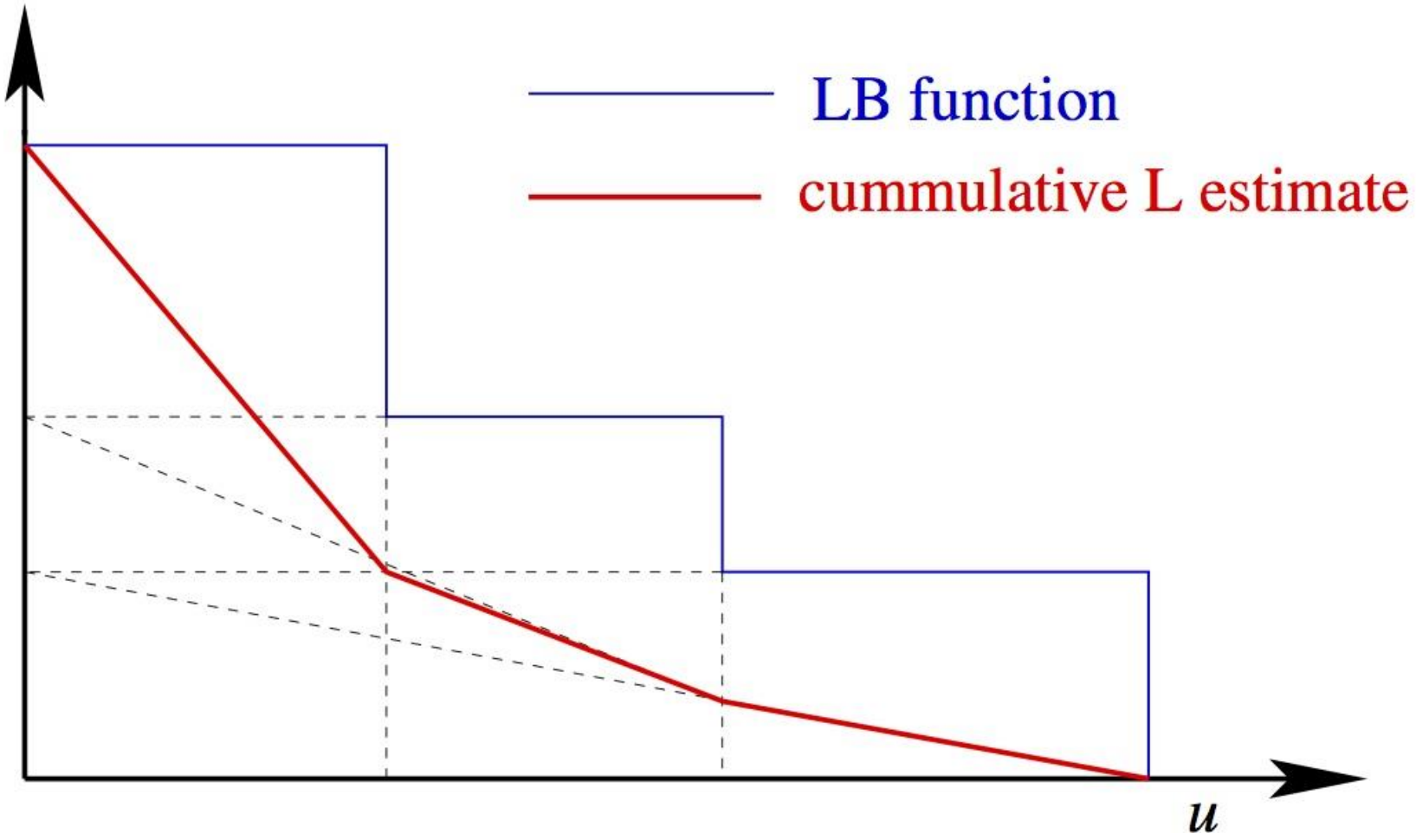
The L* estimator:

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

The U* estimator:

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

The L* estimator



Summary

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

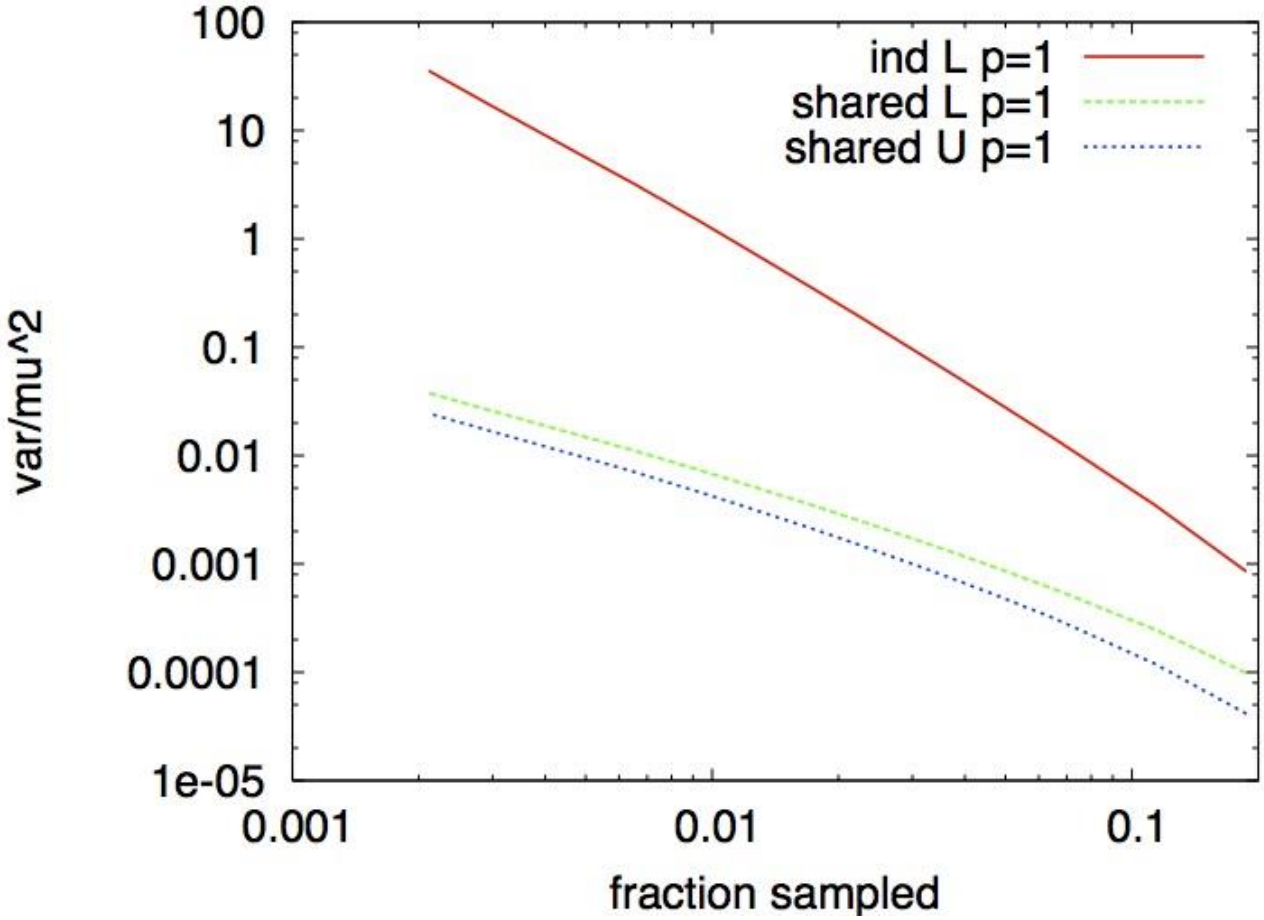
Follow-up and open problems

- 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**
- MEP with multiple seeds (independent samples)
- **Applications:**
  - Estimating Euclidean and Manhattan distances from samples [C KDD '14]
  - sketch-based similarity in social networks [CDFGGW COSN '13],
  - Timed-influence oracle [CDPW '14]

L_1 difference [C KDD14]

Independent / Coordinated PPS sampling

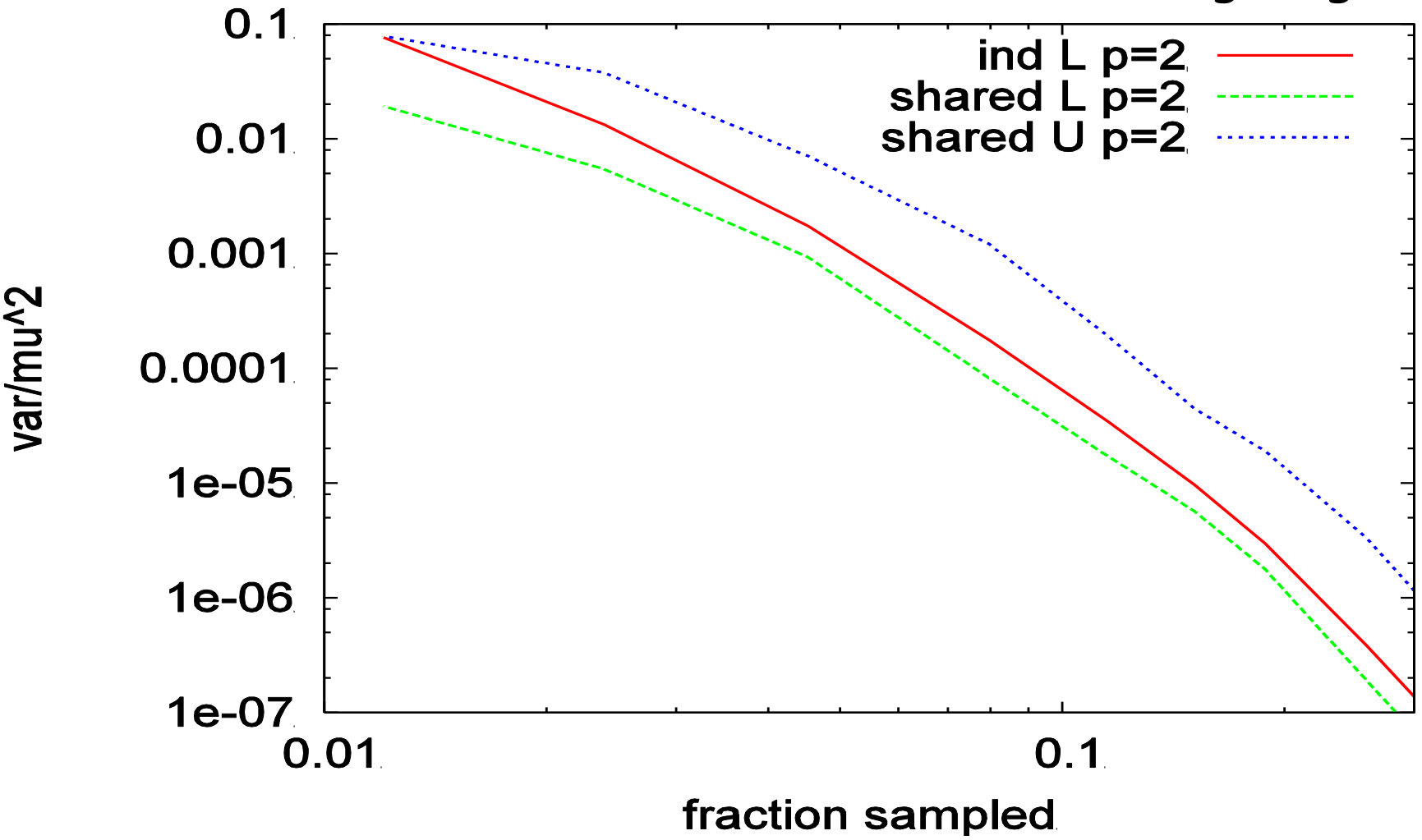
#IP flows to a destination in two time periods



L_2^2 difference [C KDD14]

Independent/Coordinated PPS sampling

Surname occurrences in 2007, 2008 books (Google ngrams)



Thank you!

Why Coordinate Samples?

- Minimize overhead in repeated surveys (also storage)

  Brewer, Early, Joice 1972; Ohlsson '98 (Statistics) …

- Can get better estimators

  Broder '97; Byers et al Tran. Networking '04; Beyer et al SIGMOD '07; Gibbons VLDB '01 ;Gibbons Tirthapurta SPAA '01; Gionis et al VLDB '99; Hadjieleftheriou et al VLDB 2009; Cohen et al '93-'13 ….

- Sometimes cheaper to compute

  Samples of neighborhoods of all nodes in a graph in linear time Cohen '93 …

Variance Competitiveness [CK13]

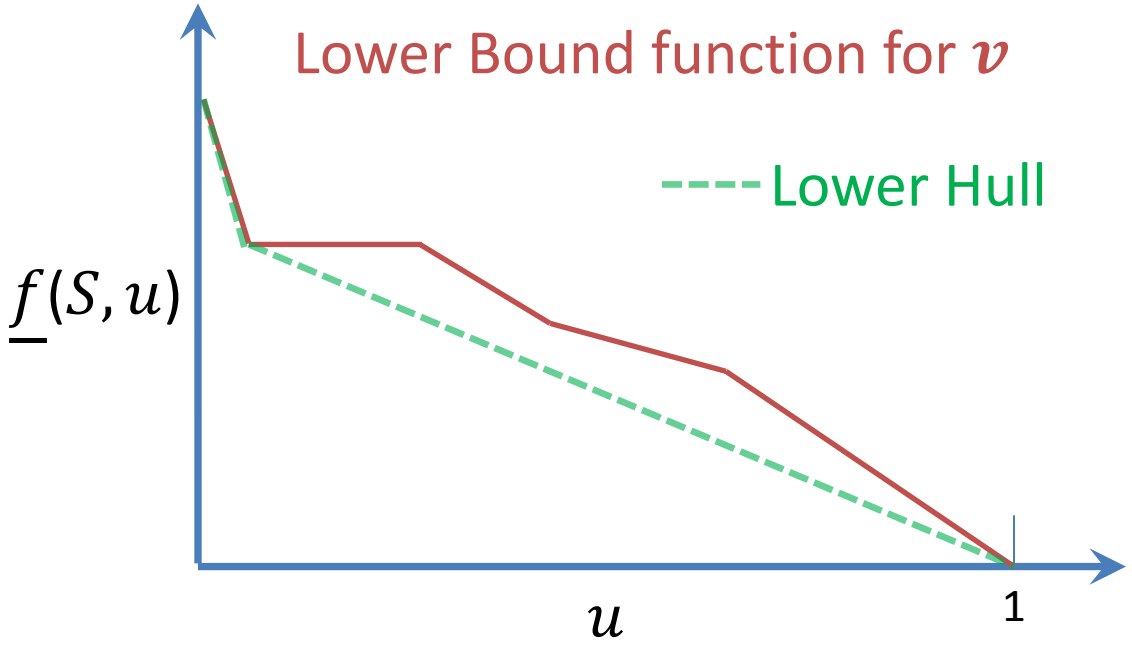
An estimator $\hat{f}(S, u)$ is ***c-competitive*** if for any data $\boldsymbol{v}$, the expectation of the square is within a factor c of the minimum possible for $\boldsymbol{v}$ (by an unbiased and nonnegative estimator).

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

$$E\left[\hat{f}^2(\mathrm{S}, \mathrm{u}) \mid \boldsymbol{v}\right] \leq c \; E\left[\hat{g}^2(\mathrm{S}, \mathrm{u}) \mid \boldsymbol{v}\right]$$

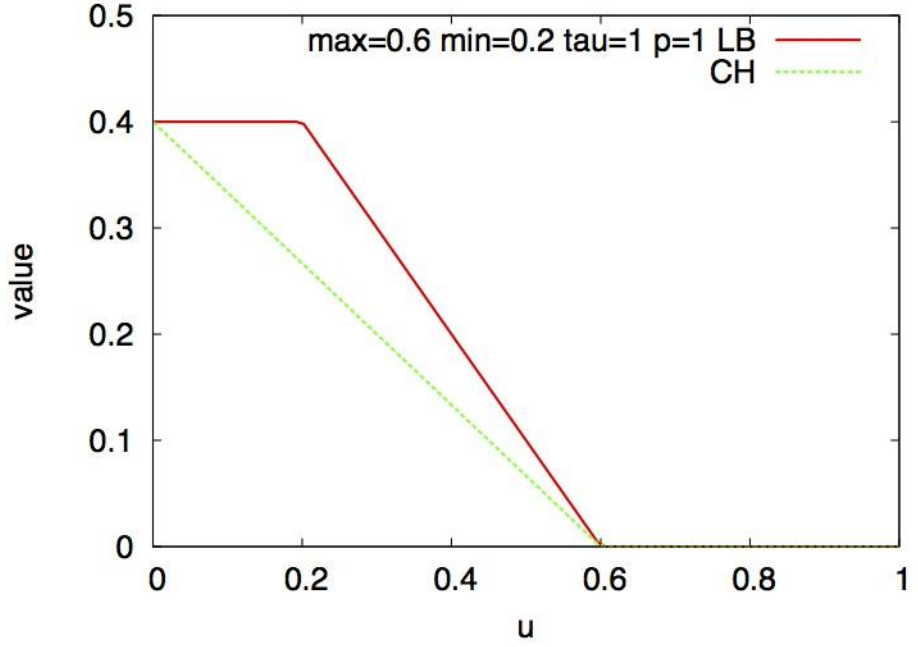
Optimal estimates $\hat{f}^{(\boldsymbol{v})}$ for data $\boldsymbol{v}$

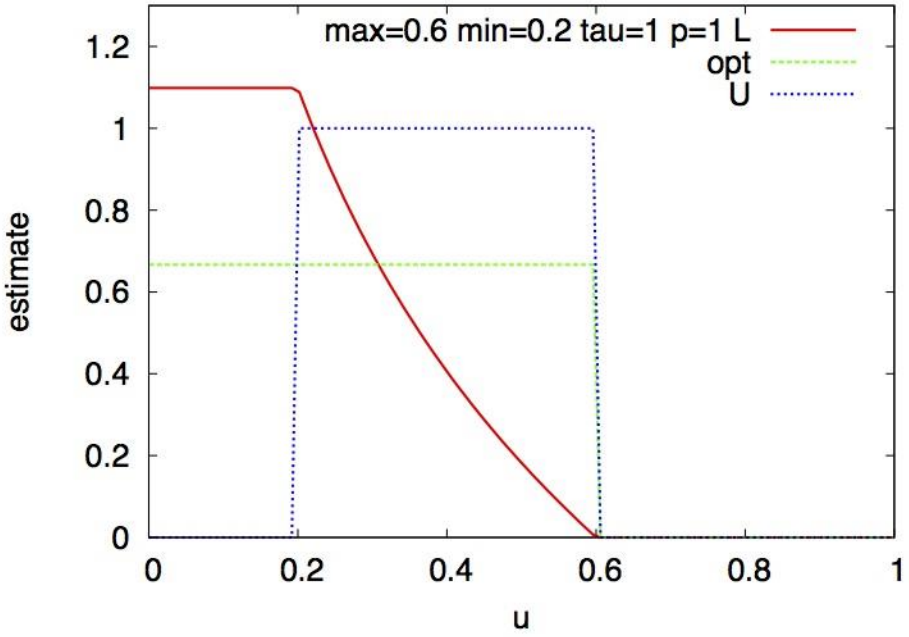
The optimal estimates $\hat{f}^{(\boldsymbol{v})}$ are the negated derivative of the lower hull of the Lower bound function.



Intuition: The lower bound tell 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.

Manhattan Distance





Euclidean Distance