What you can do with Coordinated Sampling

Edith Cohen
Microsoft Research SVC
Tel Aviv University

Haim Kaplan
Tel Aviv University

The Item-Function Estimation Problem

IFE instance $(\boldsymbol{V}, \boldsymbol{\tau}, f)$:

- Data domain $\boldsymbol{V} \subset R^r$
- Scalar $\boldsymbol{\tau}$
- A function $f: \boldsymbol{V} \geq 0$

Sampling scheme: Applied to data $\boldsymbol{v} \in \boldsymbol{V}$ to obtain a sample S

- Draw random *seed* $u \sim U[0,1]$
- Include $v_i \in S \leftrightarrow v_i \geq \boldsymbol{\tau} \cdot u$

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

- A Less general formulation is used for this talk: $\boldsymbol{\tau}$ can be different in each coordinate and we can use a general non-decreasing $\boldsymbol{\tau}(u)$

Scenario: 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 some $\boldsymbol{\tau} > 0$, independent $u(a,b)$:

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

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

Samples of multiple days

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

Example Queries: L_p difference (over selected pairs)

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

Back to the IFE problem

Many interesting queries:

L_p difference, distinct counts, quantile sums,

can be expressed as sums (or simple functions of such sums) over selected items 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 difference: $f(\boldsymbol{v}) = |v_1 - v_2|^p$

We can apply a linear (sum) estimator

$$\sum_h \hat{f}(\boldsymbol{v}^{(h)})$$

Each summand is an IFE estimator

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 …

➢ Coordination had been used for 40+ years. Many applications and independent lines of research. It is time to understand it better.

Desirable Estimator Properties

? When can we obtain an estimator $\hat{f}(S, u)$ that is:

- **Unbiased**: because bias adds up
- **Nonnegative**: because f is
- **Bounded variance** (for all $\boldsymbol{v}$)
- **Bounded** by a function of $f(\boldsymbol{v})$ (implies bounded variance)

Our Results (1)

Complete characterization in terms of $(\boldsymbol{V}, \boldsymbol{\tau}, f)$ for when the IFE instance has an estimator which is.

- **Unbiased** and **Nonnegative**
- **Unbiased**, **Nonnegative**, and has **bounded variances**
- **Unbiased**, **Nonnegative**, and **bounded**

Variance Competitiveness

What about getting a "good" estimator $\hat{f}(S,u)$?

- **Unbiased**, **Nonnegative, Bounded variance** estimators are not unique
- No ***UMVUE*** (Uniform Minimum Variance Unbiased estimator) in general.

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 $\hat{g}$,

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

Our Results (2)

Thm: For any IFE instance $(\boldsymbol{V}, \boldsymbol{\tau}, f)$ for which an unbiased, nonnegative, and bounded-variances estimator exists, we can construct an estimator that is **O(1)-competitive** (84-competitive).

- In particular, we establish the **existence** of variance competitive estimators
- The construction is fairly efficient given reasonable representation of $\boldsymbol{\tau}, f$

What is the minimum variance for data $\boldsymbol{v}$?

An important tool we use (to bound competitiveness and establish existence of bounded-variance estimator):

For data $\boldsymbol{v}$, we give an explicit construction of a "partial" estimator, $\hat{f}^{(\boldsymbol{v})}$, defined only on outcomes consistent with $\boldsymbol{v}$.

The estimates $\hat{f}^{(\boldsymbol{v})}$ minimize the variance for $\boldsymbol{v}$ under the constraint that the partial specification $\hat{f}^{(\boldsymbol{v})}$ can be completed to an estimator that is unbiased and nonnegative everywhere.

- We give the intuition for this construction.
- Turns out that $\hat{f}^{(\boldsymbol{v})}$ is unique.

The lower bound function

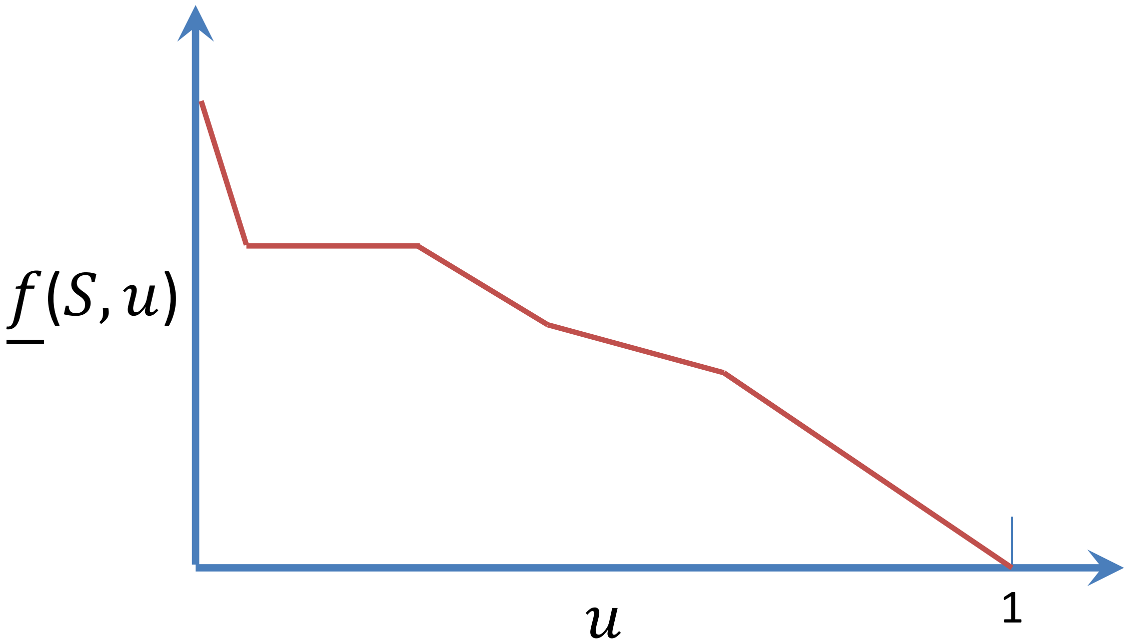
For an outcome and seed (S,u) we can look at the set of all consistent data vectors: $V^*(S,u)$

e.g. For $S = (2,*,*)$ and seed u, the set of consistent vectors includes all vectors where the second and third entries are at most $u\,\boldsymbol{\tau}$.

The lower bound $\underline{f}$(S,u) is the infimum of f on $V^*(S,u)$

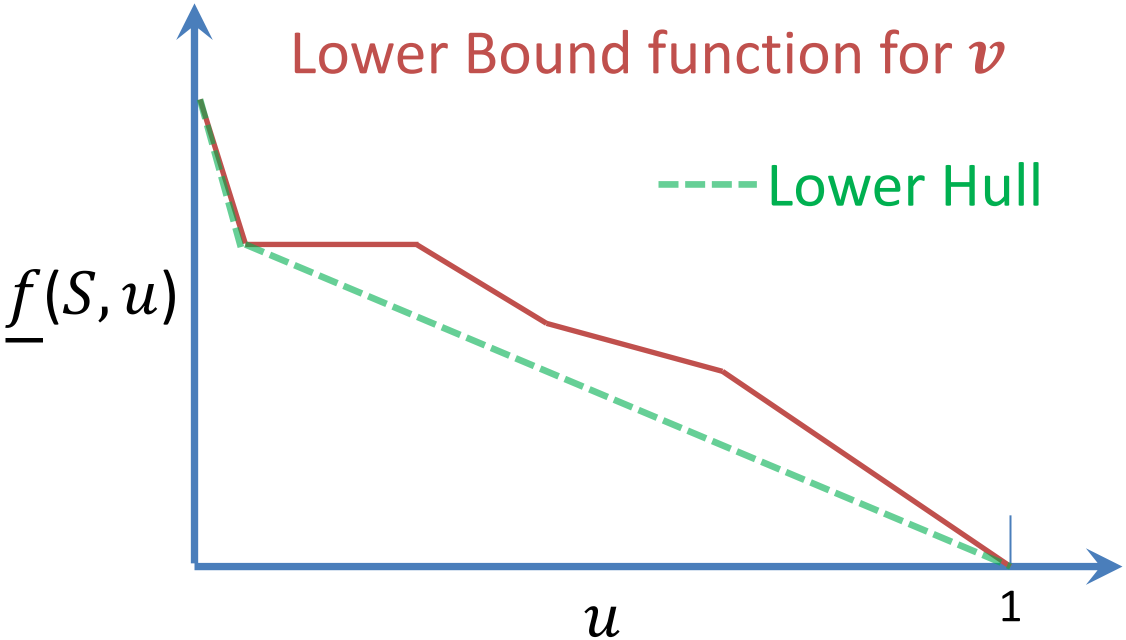
Lower bound function for data $\boldsymbol{v}$

Fix the data $\boldsymbol{v}$. Consider the lower bound $\underline{f}(S, u)$ as a function of the seed u. The lower u is, the more we know on $\boldsymbol{v}$ and hence on $f(\boldsymbol{v})$. Therefore, $\underline{f}(S, u)$ is non-decreasing



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.

Follow-up work + Open problems

- Studied range of Pareto optimal (admissible) estimators:

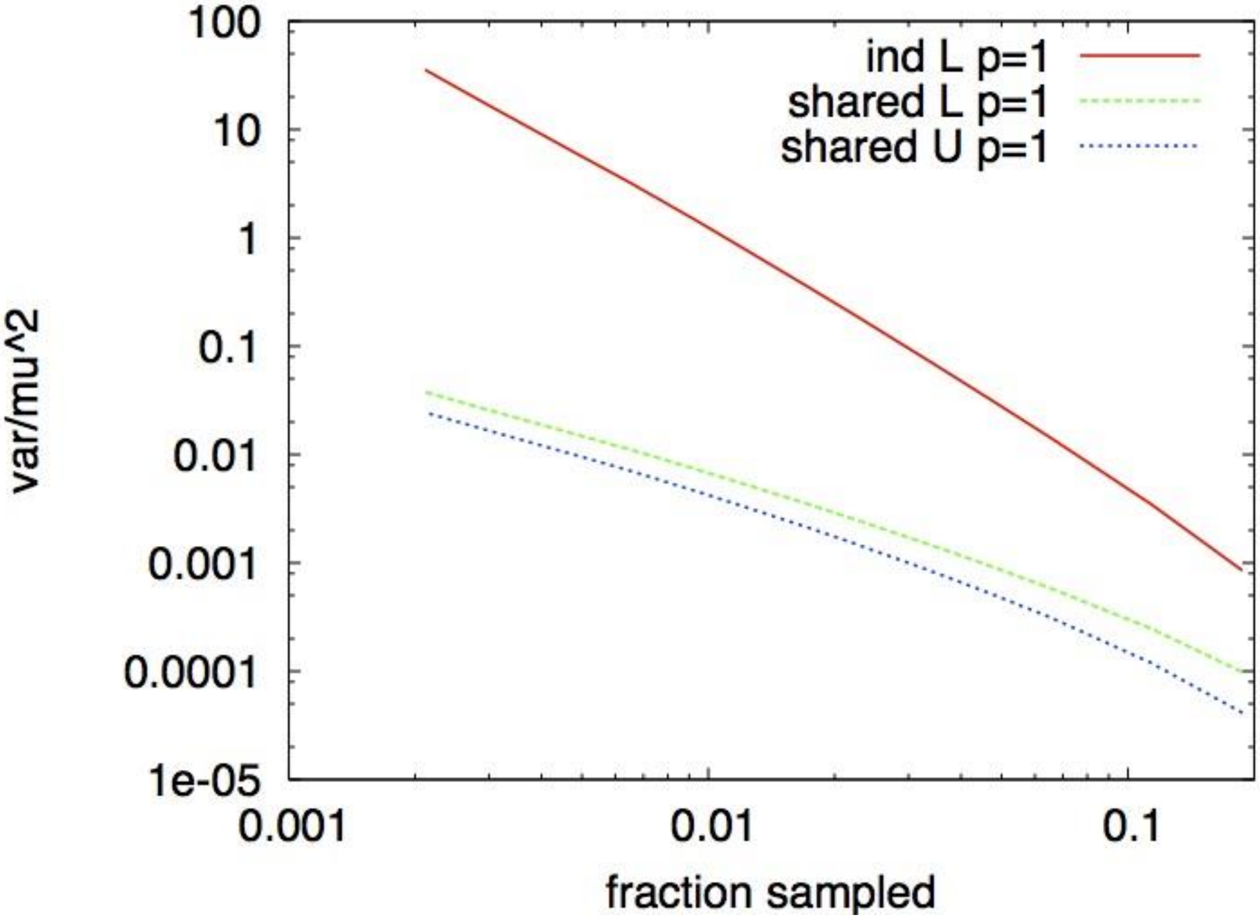
  Natural estimators: L* (lower end of range: unique monotone estimator, dominates HT) , U* (upper end of range), order optimal estimators (optimized for certain data patterns)
- Obtained tighter competitiveness bounds: L* is 4 competitive, can do 3.375 competitive, lower bound is 1.44 competitive. **Close this gap**!
- Instance-optimal competitiveness – **Give efficient construction for any IFE instance** $(\boldsymbol{V}, \boldsymbol{\tau}, \boldsymbol{f})$.
- Independent Sampling [CK PODS '11] – **A similar characterization ?**
- **Back to practice**: Difference norms on sampled data [CK '13], sketch-based similarity in social networks [CDFGGW COSN '13].

Thank you!

Estimating L_1 difference

Independent / Coordinated, pps, known seeds

destination IP addresses: #IP flows in two time periods



Estimating L_2^2 difference

Independent / Coordinated, pps, Known seeds

Surname occurrences in 2007, 2008 books (Google ngrams)