Smooth Sensitivity and Sampling

Sofya Raskhodnikova
Penn State University

Joint work with **Kobbi Nissim** (Ben Gurion University)
and **Adam Smith** (Penn State University)

Our main contributions

- *Starting point:* Global sensitivity framework [DMNS06] (Cynthia's talk)
- Two new frameworks for private data analysis
- Greatly expand the types of information that can be released

Road map

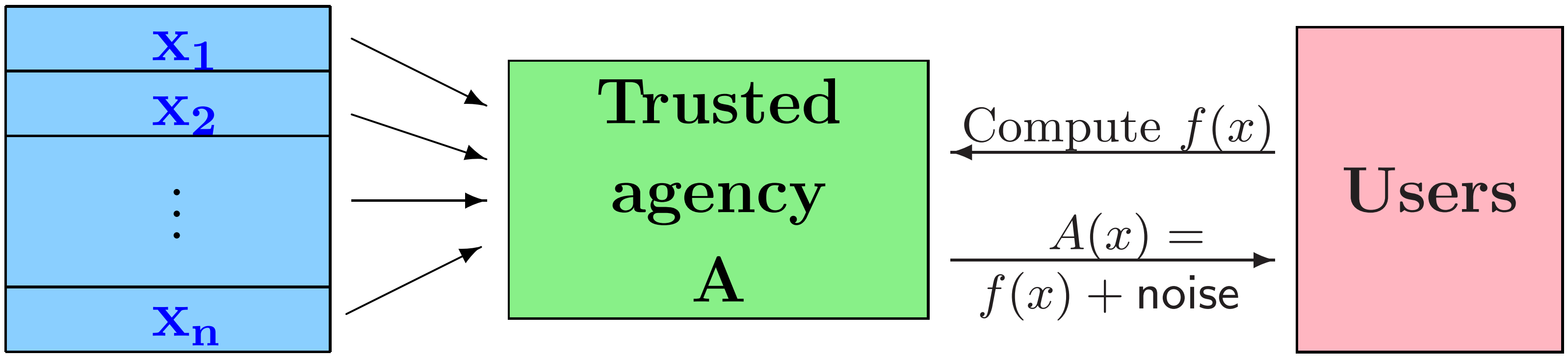
I. Introduction

- Review of global sensitivity framework [DMNS06]
- Motivation

II. Smooth sensitivity framework

III. Sample-and-aggregate framework

Model



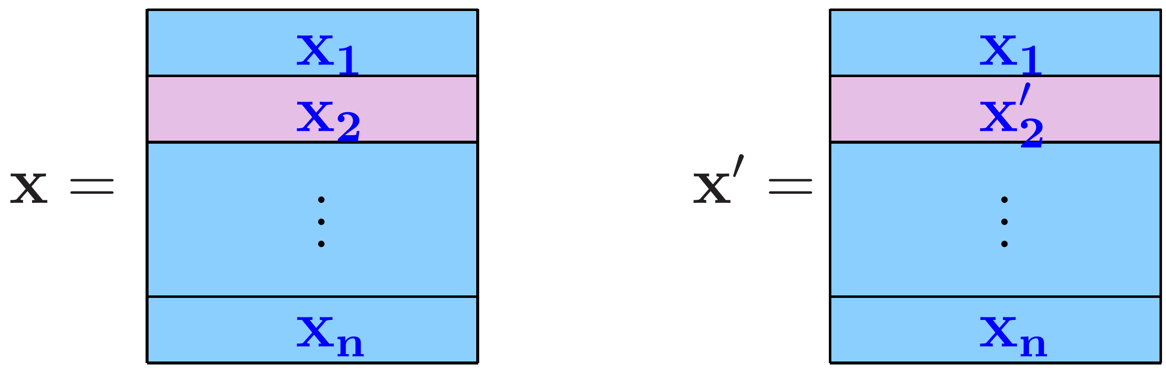
Each row is arbitrarily complex data supplied by 1 person.

For which functions f can we have:

- utility: little noise
- privacy: indistinguishability definition of [DMNS06]

Privacy as indistinguishability [DMNS06]

Two databases are *neighbors* if they differ in one row.



Privacy definition

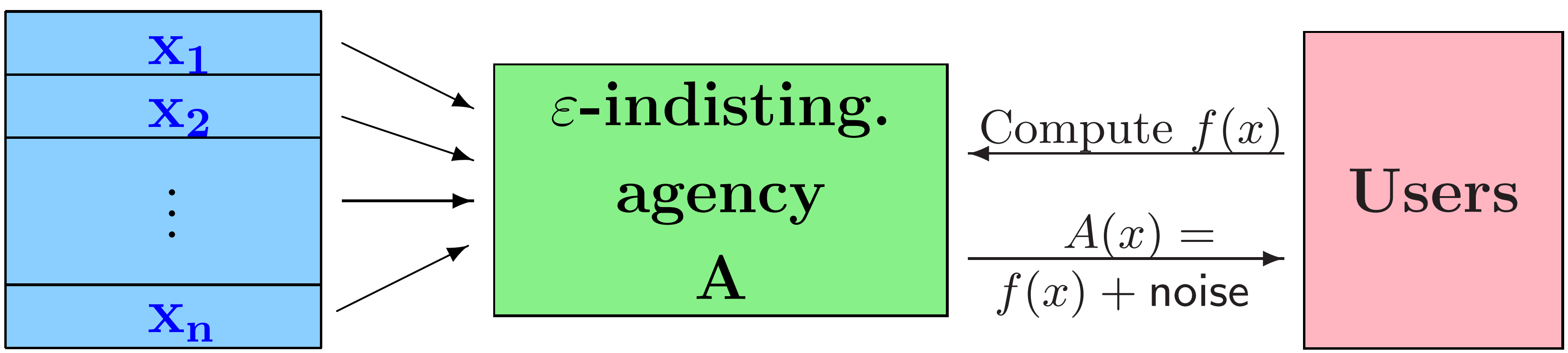
Algorithm A is ε-indistinguishable if

- *for all neighbor databases x, x'*
- *for all sets of answers S*

$$\Pr[A(x) \in S] \leq (1+\varepsilon) \cdot \Pr[A(x') \in S]$$

Privacy definition: composition

If A is ε-indistinguishable on each query,
it is εq-indistinguishable on q queries.



Global sensitivity framework [DMNS06]

Intuition: f can be released accurately when it is insensitive to individual entries $x_1, \ldots, x_n$.

Global sensitivity $\mathsf{GS}_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|$.

Example: $\mathsf{GS}_{\text{average}} = \frac{1}{n}$ if $x \in [0,1]^n$.

Theorem

If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

Instance-Based Noise

Big picture for global sensitivity framework:

- add enough noise to cover the worst case for f
- noise distribution depends only on f, not database x

Problem: for some functions that's too much noise

Instance-Based Noise

Big picture for global sensitivity framework:

- add enough noise to cover the worst case for f
- noise distribution depends only on f, not database x

Problem: for some functions that's too much noise

Example: median of $x_1, \ldots, x_n \in [0,1]$

$$x = \underbrace{0 \cdots 0}_{\frac{n-1}{2}} \, 0 \, \underbrace{1 \cdots 1}_{\frac{n-1}{2}} \qquad\qquad x' = \underbrace{0 \cdots 0}_{\frac{n-1}{2}} \, 1 \, \underbrace{1 \cdots 1}_{\frac{n-1}{2}}$$

$$\mathrm{median}(x) = 0 \qquad\qquad \mathrm{median}(x') = 1$$

$$\mathsf{GS}_{\mathrm{median}} = 1$$

- Noise magnitude: $\frac{1}{\varepsilon}$.

Instance-Based Noise

Big picture for global sensitivity framework:

- add enough noise to cover the worst case for f
- noise distribution depends only on f, not database x

Problem: for some functions that's too much noise

Example: median of $x_1, \ldots, x_n \in [0,1]$

$$x = \underbrace{0 \cdots 0}_{\frac{n-1}{2}} 0 \underbrace{1 \cdots 1}_{\frac{n-1}{2}} \qquad x' = \underbrace{0 \cdots 0}_{\frac{n-1}{2}} 1 \underbrace{1 \cdots 1}_{\frac{n-1}{2}}$$

$$\text{median}(x) = 0 \qquad \text{median}(x') = 1$$

$$\mathsf{GS}_{\text{median}} = 1$$

- Noise magnitude: $\frac{1}{\varepsilon}$.

Our goal: noise tuned to database x

Road map

I. Introduction

- Review of global sensitivity framework [DMNS06]
- Motivation

II. Smooth sensitivity framework

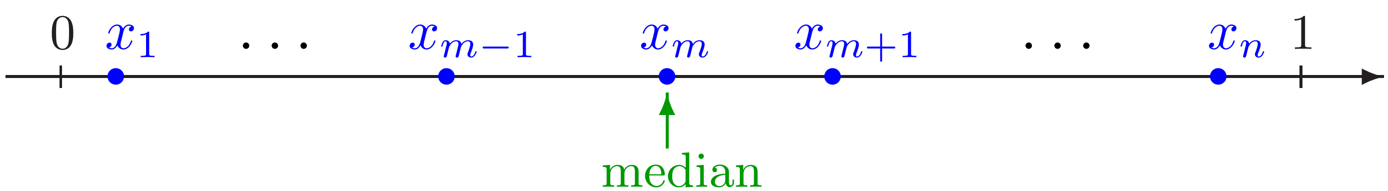
III. Sample-and-aggregate framework

Local sensitivity

Local sensitivity $\mathsf{LS}_f(x) = \max\limits_{x'\text{: neighbor of } x} \|f(x) - f(x')\|$

Reminder: $\mathsf{GS}_f = \max\limits_{x} \mathsf{LS}_f(x)$

Example: median for $0 \le x_1 \le \cdots \le x_n \le 1$, odd n



$$\mathsf{LS}_{\text{median}}(x) = \max(x_m - x_{m-1}, x_{m+1} - x_m)$$

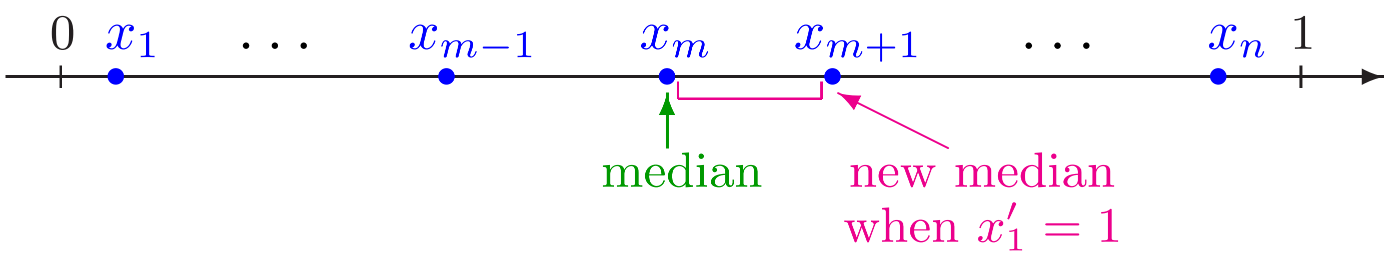
Goal: Release $f(x)$ with less noise when $\mathsf{LS}_f(x)$ is lower.

Local sensitivity

Local sensitivity $\mathsf{LS}_f(x) = \max_{x'\text{: neighbor of } x} \|f(x) - f(x')\|$

Reminder: $\mathsf{GS}_f = \max_x \mathsf{LS}_f(x)$

Example: median for $0 \le x_1 \le \cdots \le x_n \le 1$, odd n



$$\mathsf{LS}_{\text{median}}(x) = \max(x_m - x_{m-1}, x_{m+1} - x_m)$$

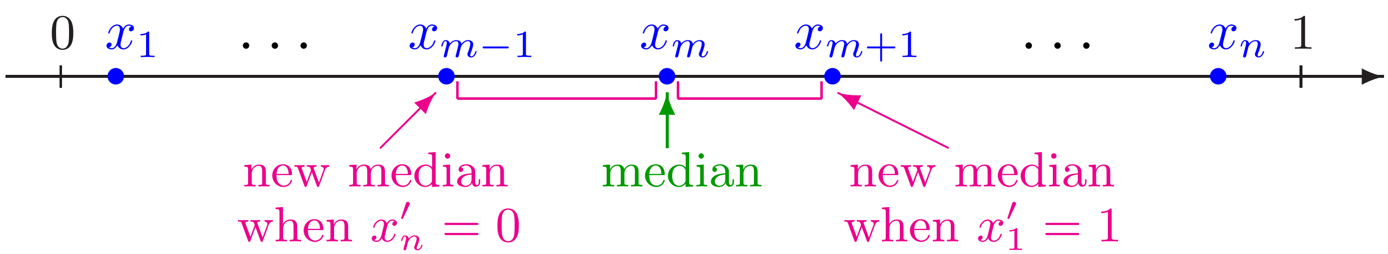
Goal: Release $f(x)$ with less noise when $\mathsf{LS}_f(x)$ is lower.

Local sensitivity

Local sensitivity $\mathsf{LS}_f(x) = \max_{x'\text{: neighbor of } x} \|f(x) - f(x')\|$

Reminder: $\mathsf{GS}_f = \max_x \mathsf{LS}_f(x)$

Example: median for $0 \le x_1 \le \cdots \le x_n \le 1$, odd n



$$\mathsf{LS}_{\text{median}}(x) = \max(x_m - x_{m-1}, x_{m+1} - x_m)$$

Goal: Release $f(x)$ with less noise when $\mathsf{LS}_f(x)$ is lower.

Instance-based noise: first attempt

Noise magnitude proportional to $\mathsf{LS}_f(x)$ instead of GS_f?

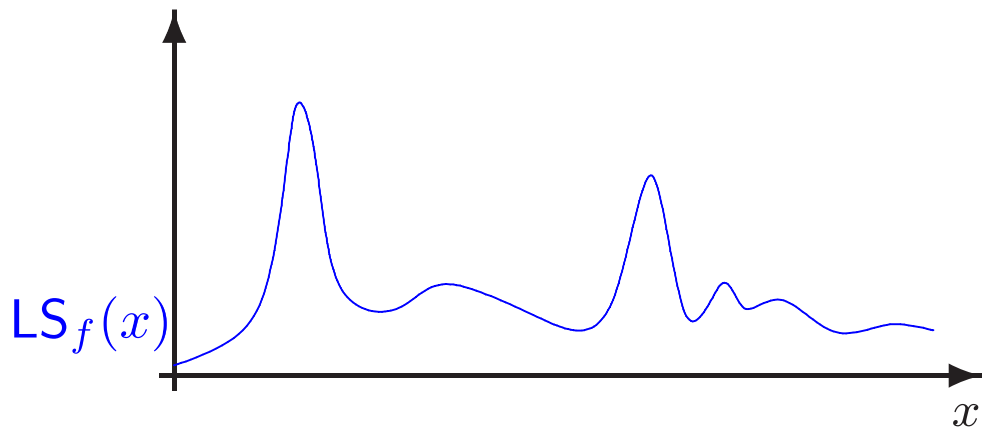
No! Noise magnitude reveals information.

Lesson: Noise magnitude must be an insensitive function.

Smooth bounds on local sensitivity

Design sensitivity function $S(x)$

- $S(x)$ is an ε-smooth upper bound on $\mathsf{LS}_f(x)$ if:
  - for all x: $S(x) \geq \mathsf{LS}_f(x)$
  - for all neighbors x, x' : $S(x) \leq e^{\varepsilon} S(x')$



Theorem

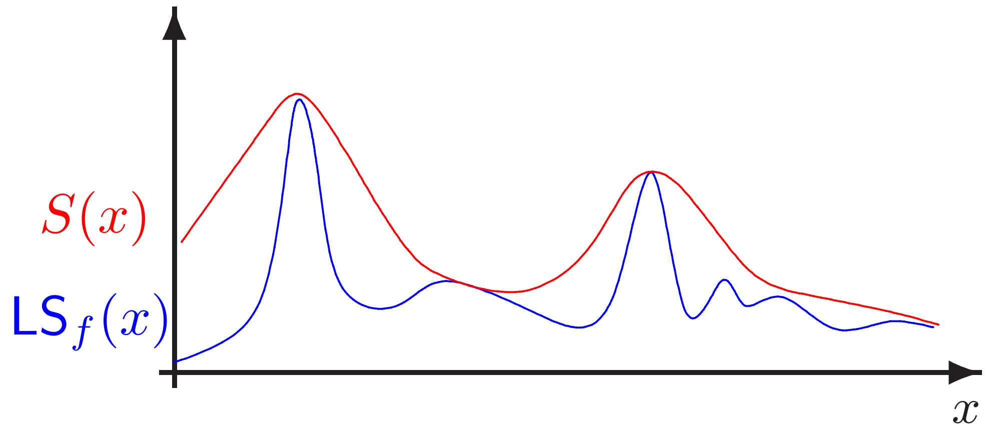
If $A(x) = f(x) + \mathsf{noise}\left(\frac{S(x)}{\varepsilon}\right)$ *then* A *is* ε'*-indistinguishable.*

Example: GS_f is always a smooth bound on $\mathsf{LS}_f(x)$

Smooth bounds on local sensitivity

Design sensitivity function $S(x)$

- $S(x)$ is an ε-smooth upper bound on $\mathsf{LS}_f(x)$ if:
  - for all x: $S(x) \geq \mathsf{LS}_f(x)$
  - for all neighbors x, x' : $S(x) \leq e^{\varepsilon} S(x')$



Theorem

If $A(x) = f(x) + \mathsf{noise}\left(\frac{S(x)}{\varepsilon}\right)$ *then* A *is* ε'*-indistinguishable.*

Example: GS_f is always a smooth bound on $\mathsf{LS}_f(x)$

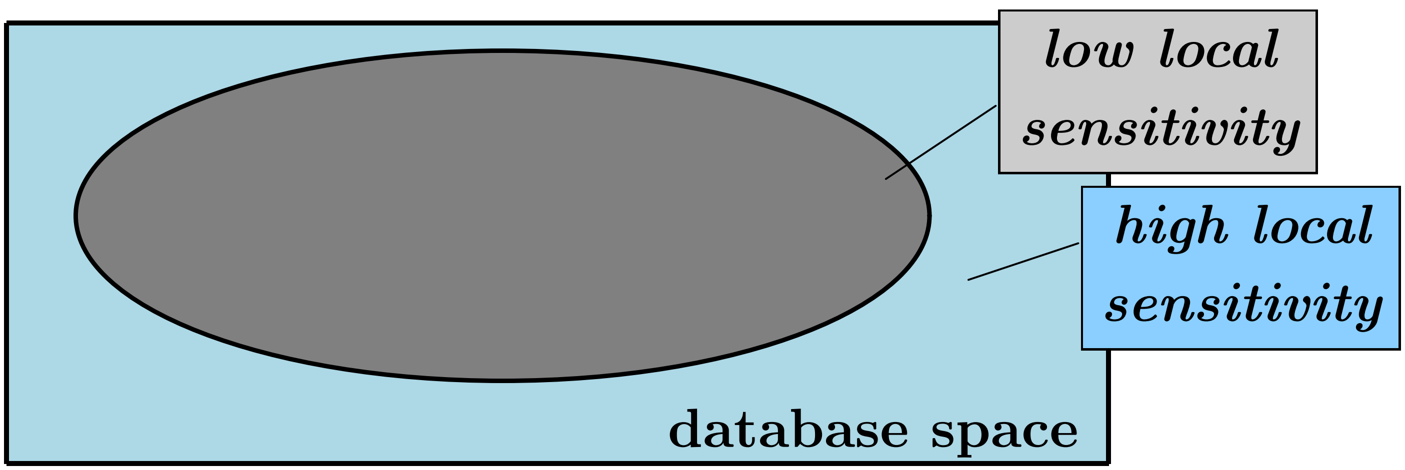
Smooth Sensitivity

Smooth sensitivity $\mathsf{S}^*_f(x) = \max_y \left(\mathsf{LS}_f(y) e^{-\varepsilon \cdot \mathsf{dist}(x,y)}\right)$

Lemma

For every ε-smooth bound S: $\mathsf{S}^*_f(x) \le S(x)$ *for all x.*

Intuition: little noise when **far** from sensitive instances



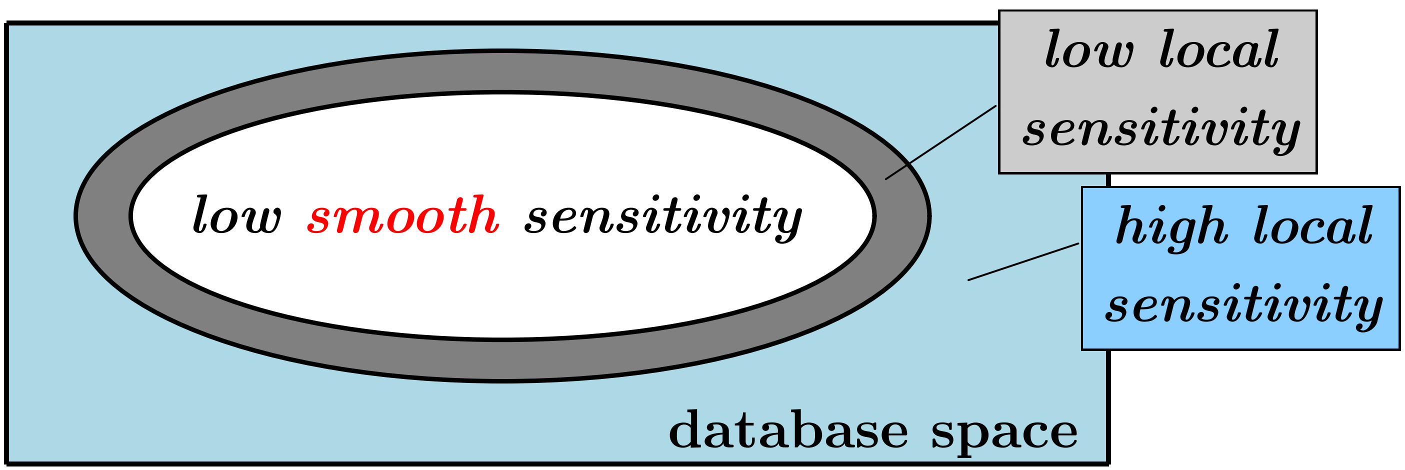
Smooth Sensitivity

Smooth sensitivity $\mathsf{S}^*_f(x) = \max_y \left(\mathsf{LS}_f(y) e^{-\varepsilon \cdot \mathsf{dist}(x,y)}\right)$

Lemma

For every ε-smooth bound S: $\mathsf{S}^*_f(x) \leq S(x)$ *for all x.*

Intuition: little noise when **far** from sensitive instances



Computing smooth sensitivity

Example functions with computable smooth sensitivity

- *Median* & *minimum* of numbers in a bounded interval
- *MST cost* when weights are bounded
- *Number of triangles* in a graph

Approximating smooth sensitivity

- only smooth upper bounds on LS are meaningful
- simple generic methods for smooth approximations
  - work for *median* and *1-median in* L_1^d

Road map

I. Introduction

- Review of global sensitivity framework [DMNS06]
- Motivation

II. Smooth sensitivity framework

III. Sample-and-aggregate framework

New goal

- Smooth sensitivity framework requires understanding combinatorial structure of f

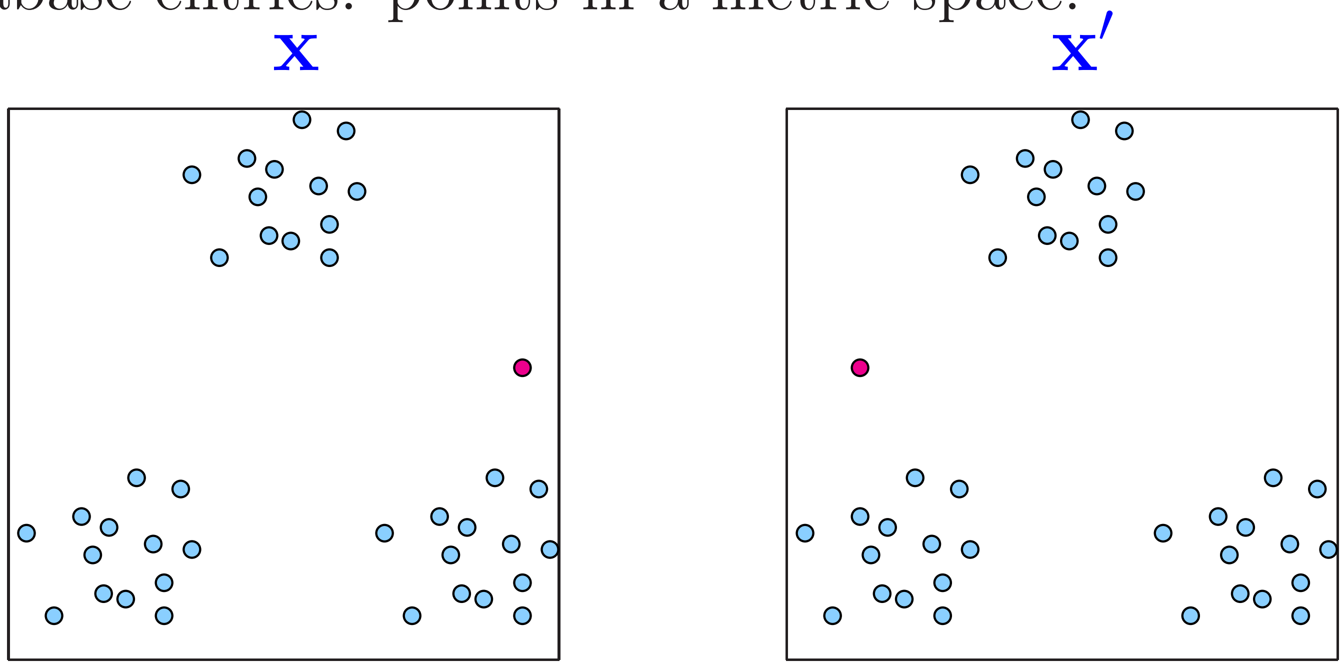
  – hard in general

- Goal: an automatable transformation from an arbitrary f into an ε-indistinguishable A

  – $A(x) \approx f(x)$ for "good" instances x

Example: cluster centers

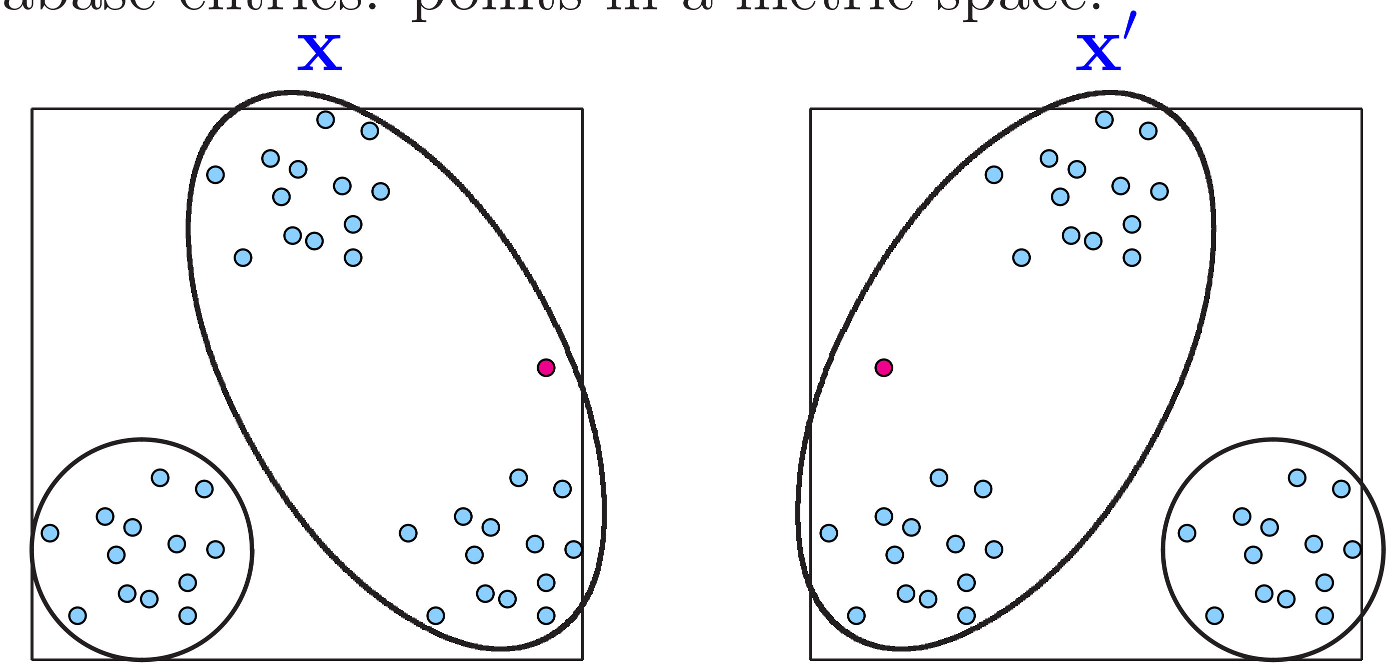
Database entries: points in a metric space.



- Comparing sets of centers: Earthmover-like metric
- Global sensitivity of cluster centers is roughly the diameter of the space. But intuitively, if clustering is "good", cluster centers should be insensitive.
- No efficient approximation for smooth sensitivity

Example: cluster centers

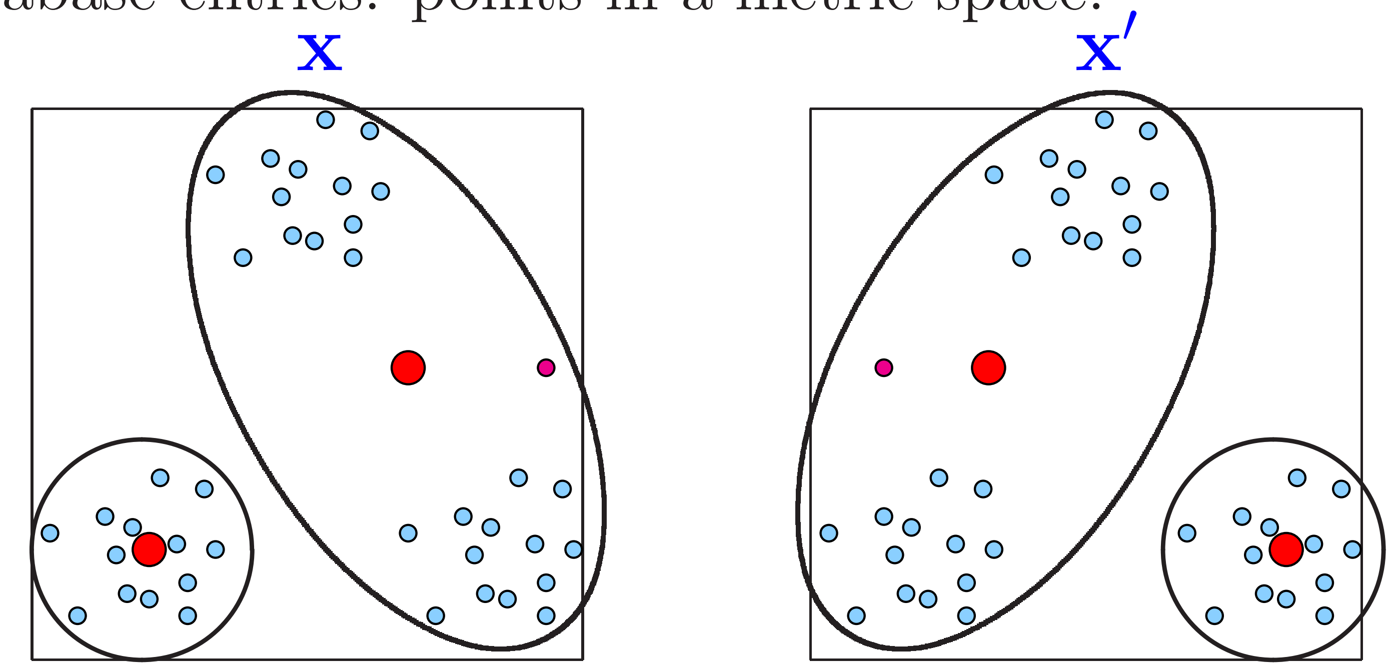
Database entries: points in a metric space.



- Comparing sets of centers: Earthmover-like metric
- Global sensitivity of cluster centers is roughly the diameter of the space. But intuitively, if clustering is "good", cluster centers should be insensitive.
- No efficient approximation for smooth sensitivity

Example: cluster centers

Database entries: points in a metric space.

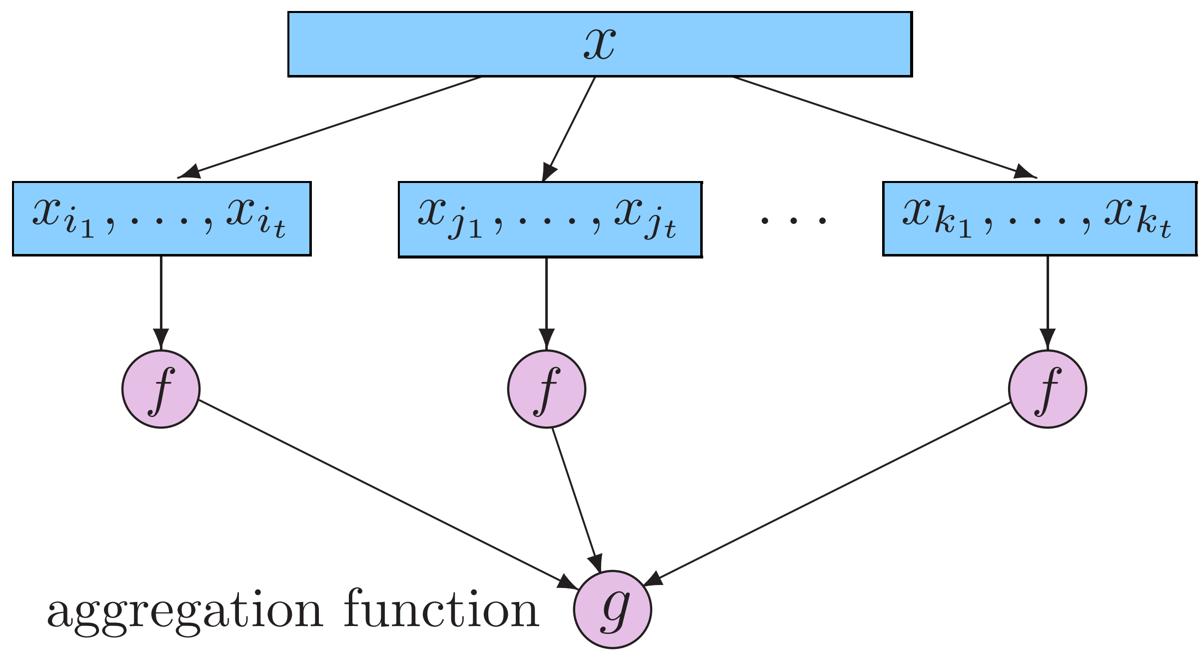


- Comparing sets of centers: Earthmover-like metric
- Global sensitivity of cluster centers is roughly the diameter of the space. But intuitively, if clustering is "good", cluster centers should be insensitive.
- No efficient approximation for smooth sensitivity

Sample-and-aggregate framework

Intuition: Replace f with a less sensitive function $\tilde{f}$.

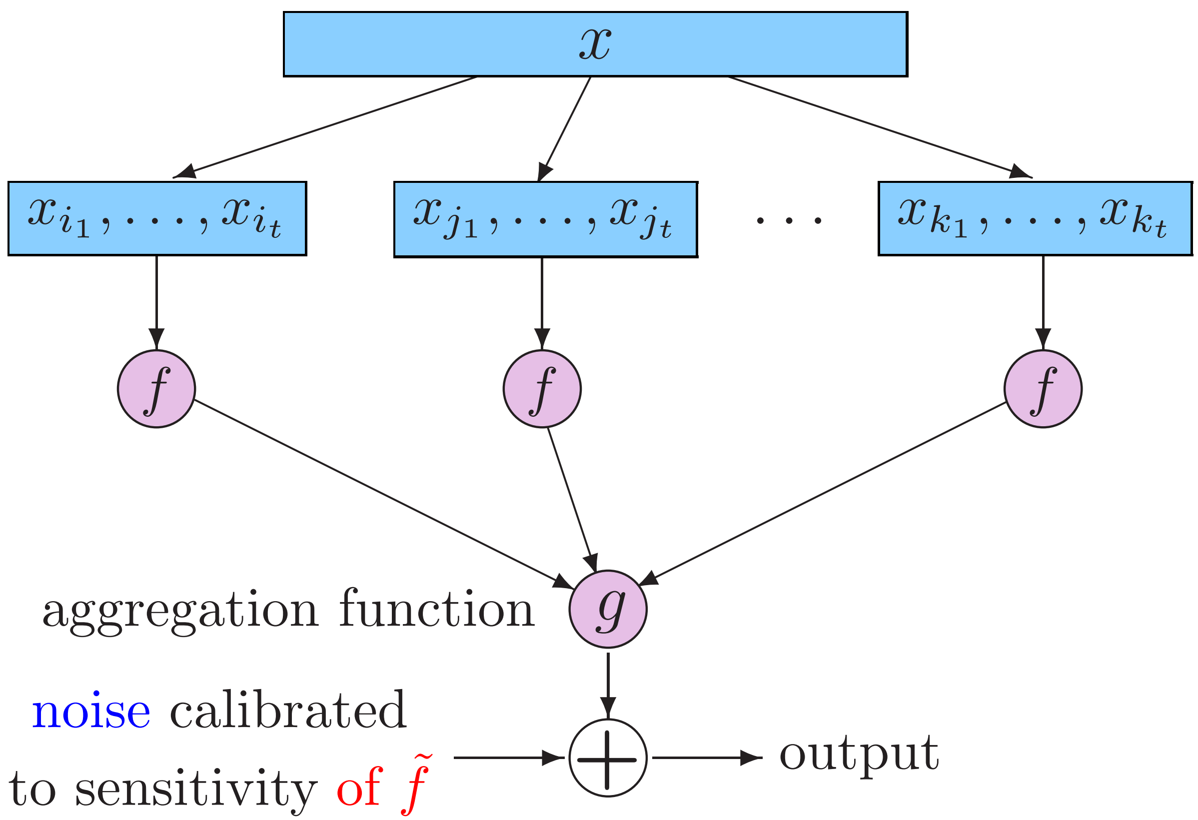
$$\tilde{f}(x) = g(f(sample_1), f(sample_2), \ldots, f(sample_s))$$



Sample-and-aggregate framework

Intuition: Replace f with a less sensitive function $\tilde{f}$.

$$\tilde{f}(x) = g(f(sample_1), f(sample_2), \ldots, f(sample_s))$$



Good aggregation functions

- average
  - works for L_1 and L_2
- center of attention
  - the center of a smallest ball containing a strict majority of input points
  - works for arbitrary metrics (in particular, for Earthmover)
  - gives lower noise for L_1 and L_2

Sample-and-aggregate results

Theorem

If f can be approximated on x

from small samples

then f can be released with little noise

Sample-and-aggregate results

Theorem

If f can be approximated on x within distance r

from small samples of size $n^{1-\delta}$

then f can be released with little noise $\approx \frac{r}{\varepsilon} + negl(n)$

Sample-and-aggregate results

Theorem

If f can be approximated on x within distance r

from small samples of size $n^{1-\delta}$

then f can be released with little noise $\approx \frac{r}{\varepsilon} + negl(n)$

- Works in all "interesting" metric spaces
- Example applications
  - k-means cluster centers (if data is separated a.k.a. [Ostrovsky Rabani Schulman Swamy 06])
  - fitting mixtures of Gaussians (if data is i.i.d., using [Vempala Wang 04, Achlioptas McSherry 05])
  - PAC concepts (Adam Smith's talk)

Road map

I. Introduction

- Review of global sensitivity framework [DMNS06]
- Motivation

II. Smooth sensitivity framework

III. Sample-and-aggregate framework

Conclusion: fundamental question

Which computations are not too sensitive to individual inputs?

Which functions f admit ε-indistinguishable approximation A?