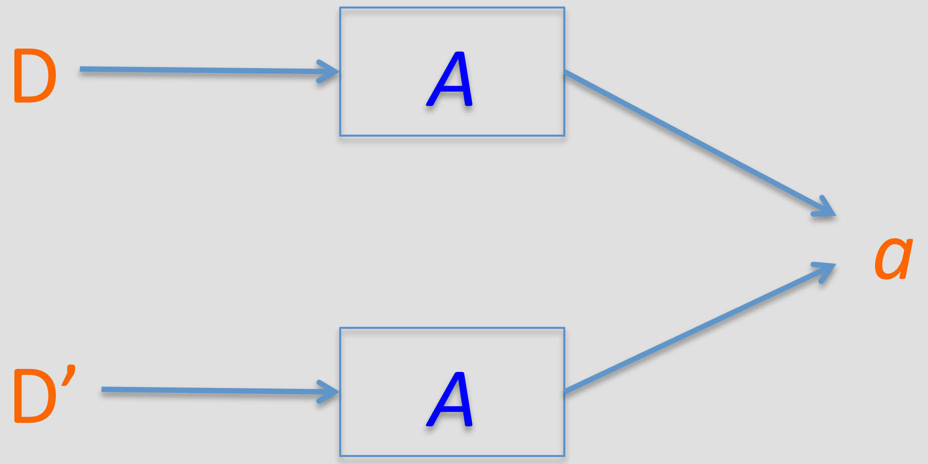
Adaptive Differentially Private Histogram of Low-Dimensional Data

Chengfang Fang Ee-Chien Chang
School of Computing
National University of Singapore

Background: Differential Privacy

A mechanism *A* achieves (Bounded) ε-Differential Privacy, if



for any published *a* and any pair of "neighbouring" datasets D and D',

$$e^{-\varepsilon} \leq \frac{\Pr[\ A(D) = a]}{\Pr[\ A(D') = a]} \leq e^{\varepsilon}$$

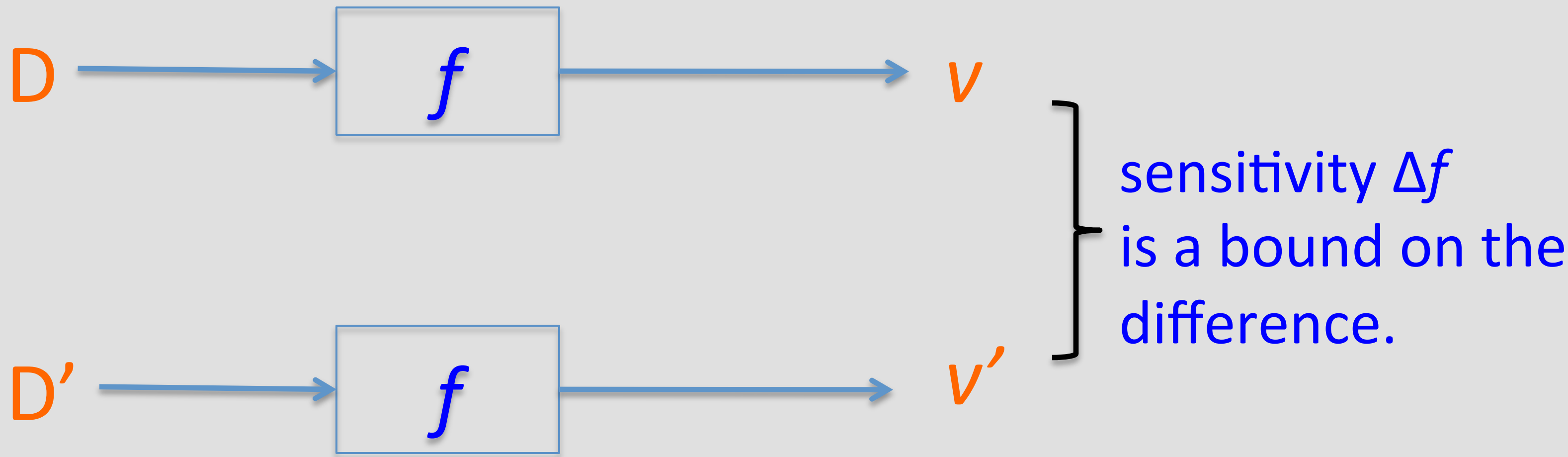
“Bounded” diff. privacy

D and D’ are neighbours

iff

D’ can be obtained from D by replacing one element.

Background: Sensitivity



The sensitivity of $f: \mathcal{D} \rightarrow R^n$, denoted as Δf, is defined as:

$$\Delta f = \max_{D,D'} |f(D) - f(D')|_1$$

where max is taken over all neighbouring D,D'.

Background: Sensitivity → diff. priv. [Dwork06]

If sensitivity of a function f is Δf then the mechanism A

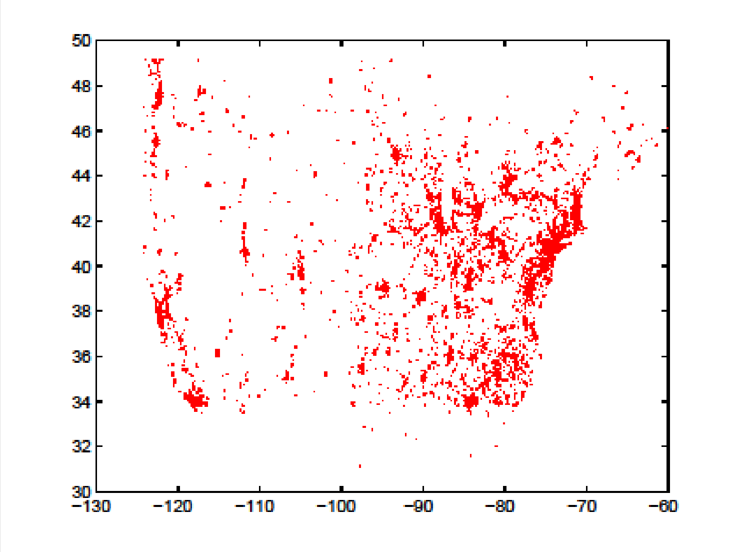
$$A(D) = f(D) + LAP(\Delta f / \varepsilon)$$

achieves ε-differential privacy.

Problem: illustrating examples

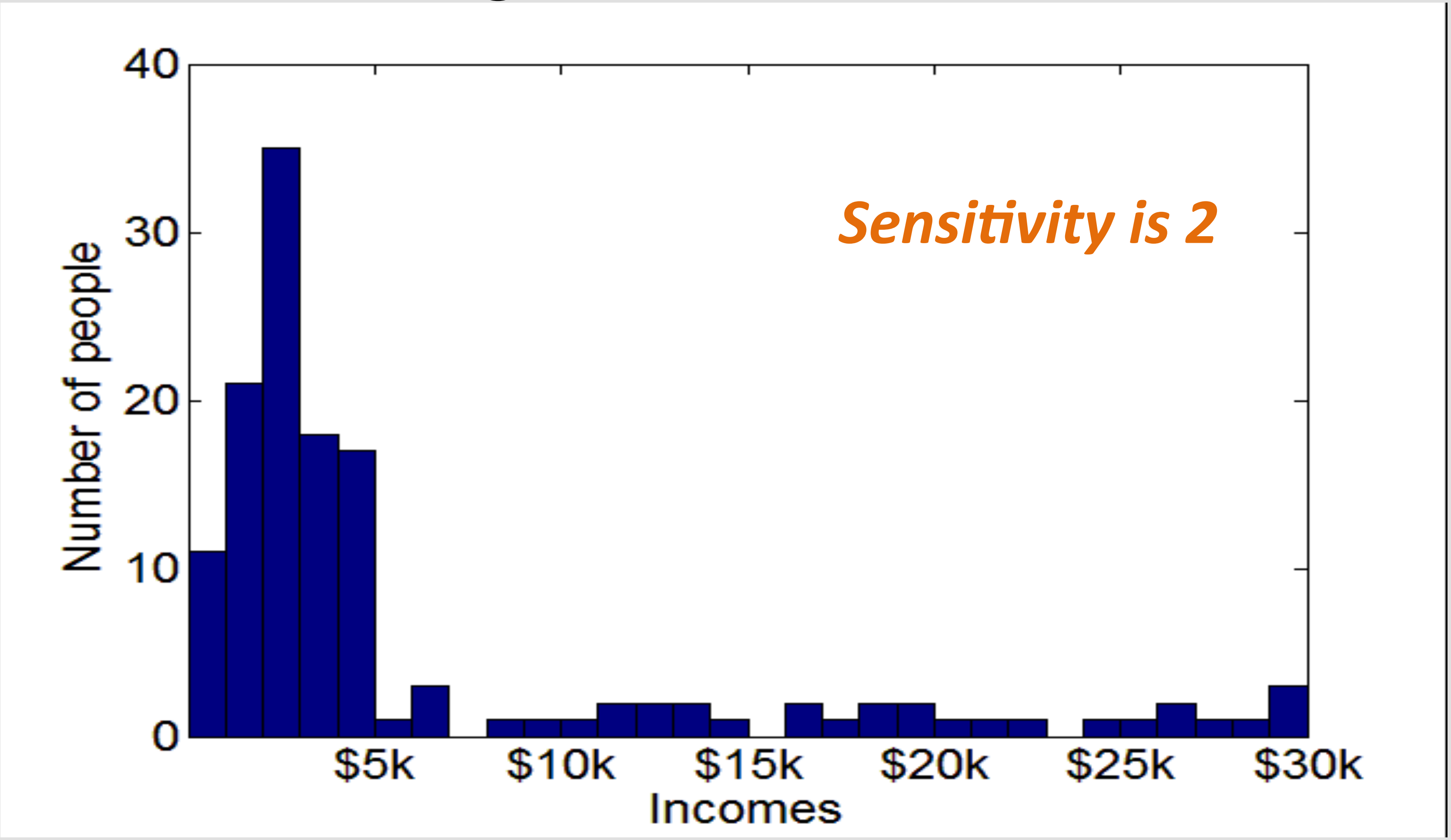
- We want to publish the "distribution" of a dataset D in a differentially private manner.
  - e.g. incomes of a group of taxpayers,
    D={ $10031, $8931, $3001, $21530, …., $32320 }
  - e.g Locations of individuals

D=



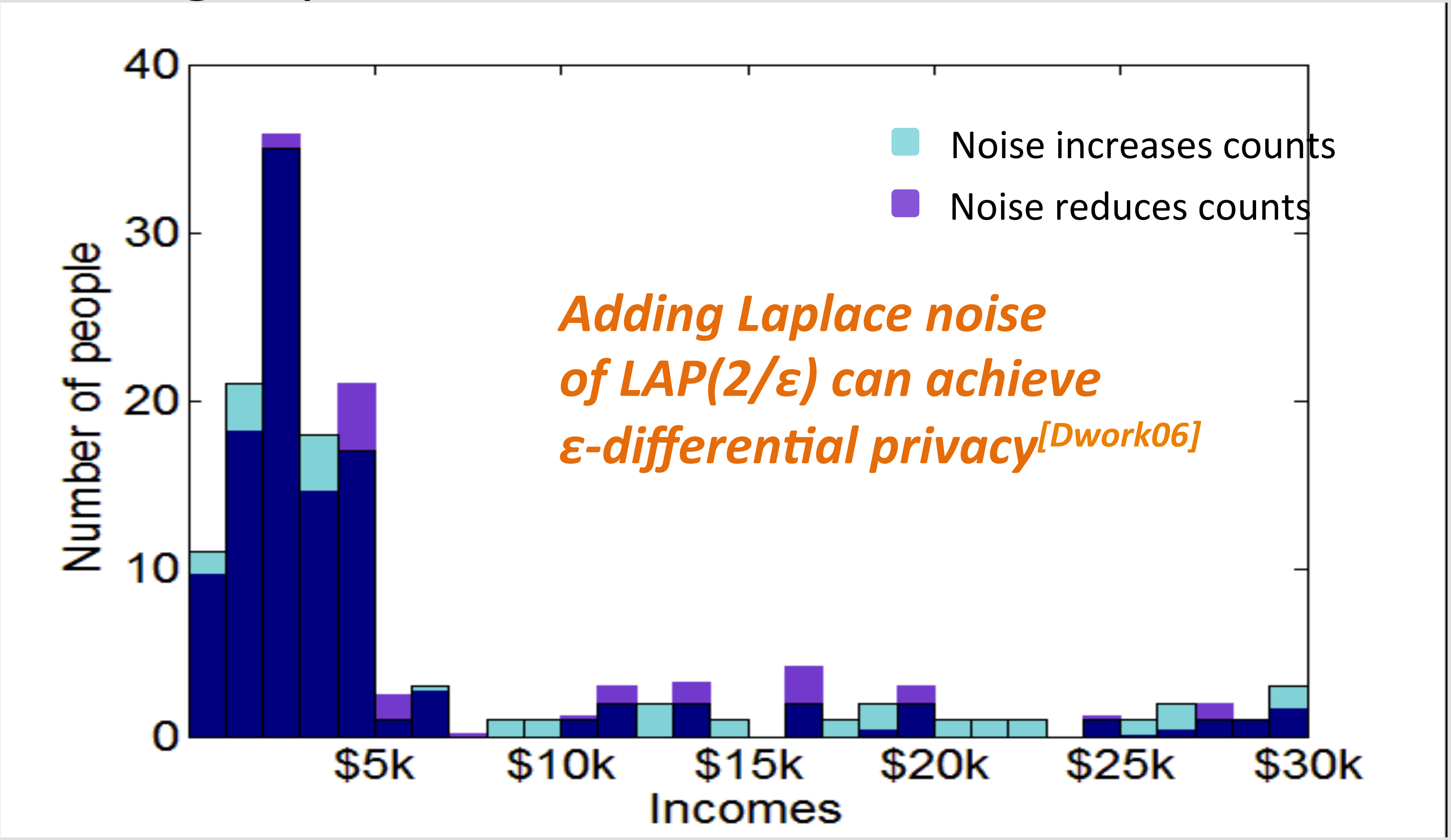
Existing approach: Equi-width Histogram

The actual histogram with 30 bins.



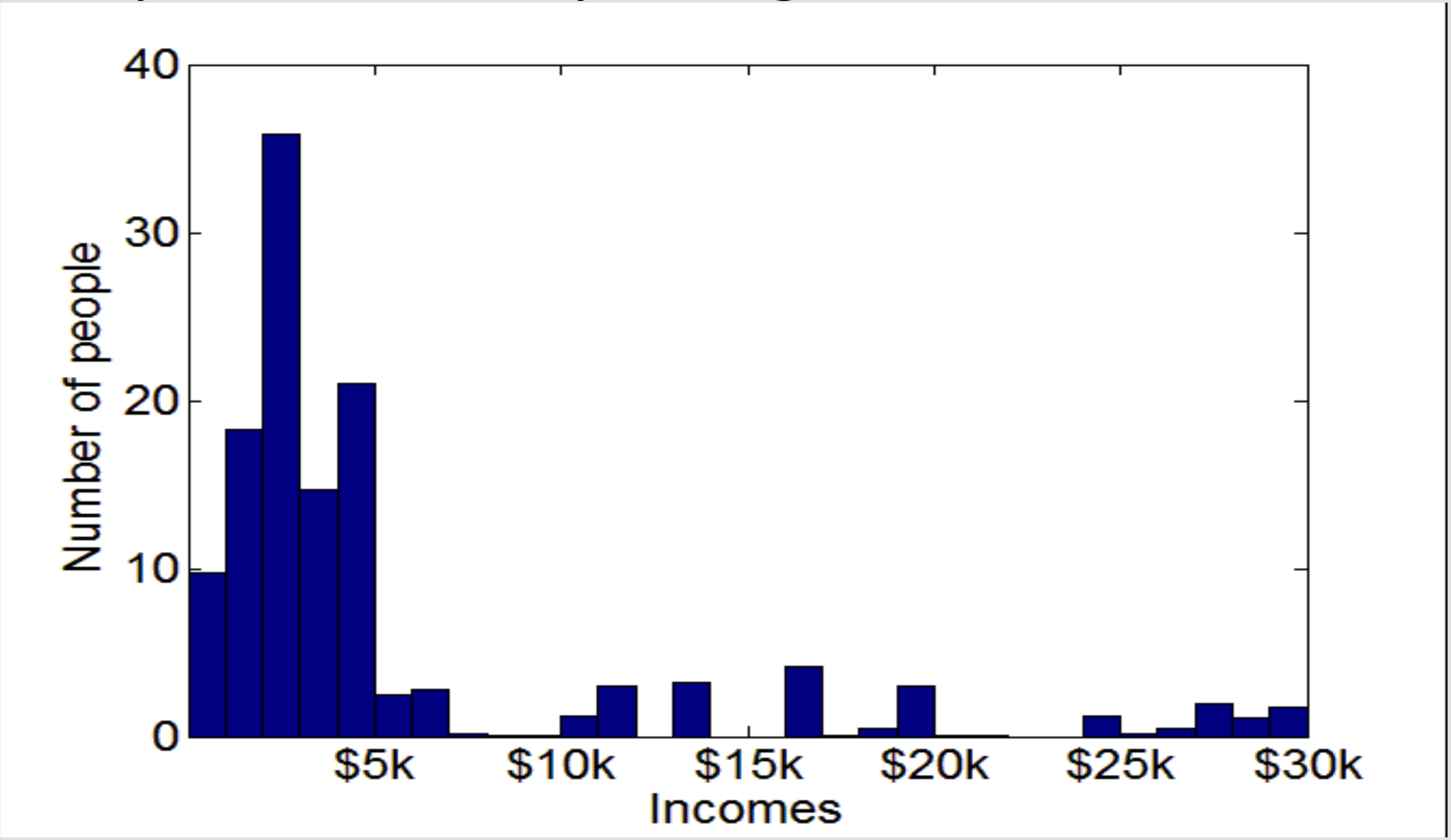
Existing approach: Equi-width Histogram

Adding Laplace noise to the counts.

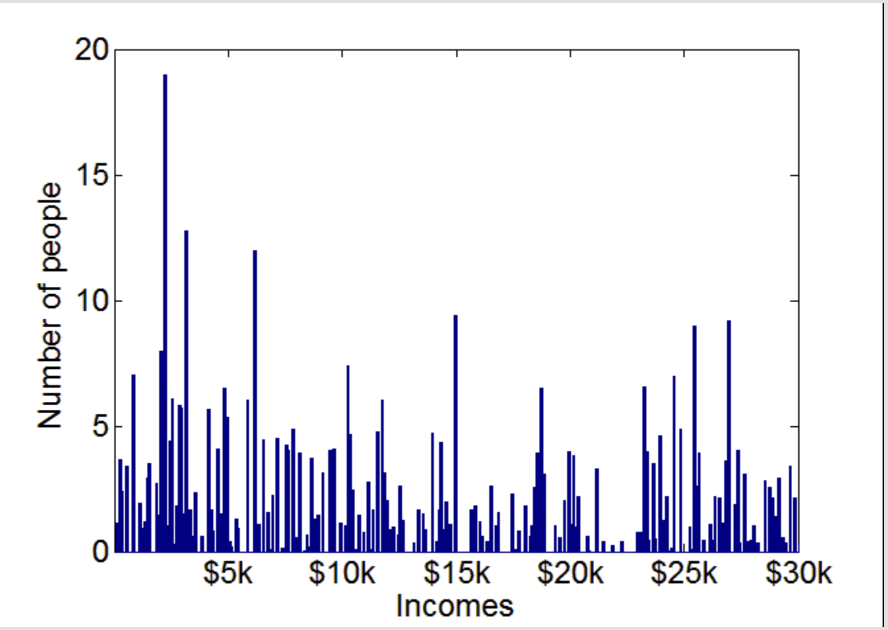


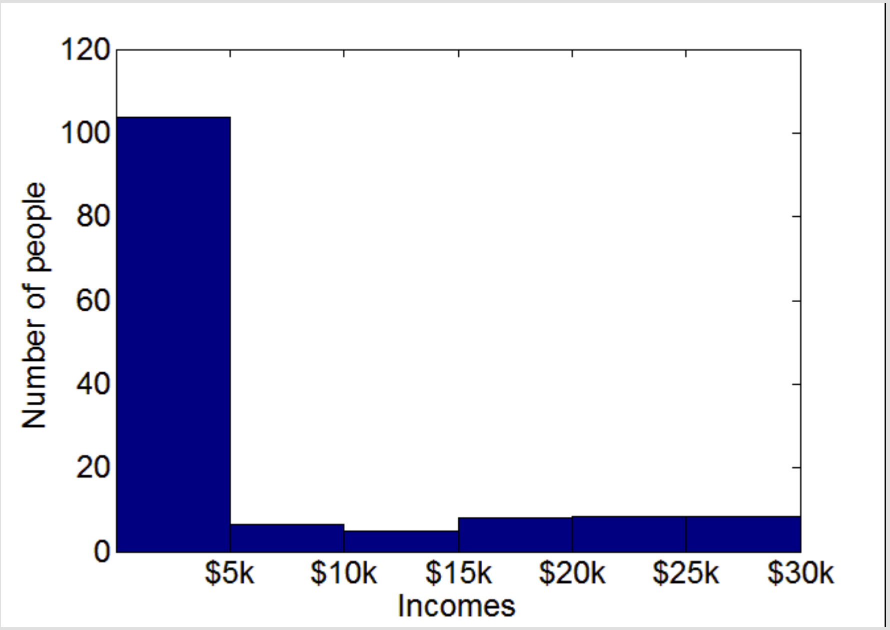
Existing approach: Equi-width Histogram

The published noisy histogram.



Problem with Equi-Width Histogram





Small bin-width:

Incur too much noise.

Large bin-width:

Lost detail information.

Enhancements and variations

- Wavelet-based: Publishing a series of histograms [Xiao10, Hay10,Chan11].
- Exploit dependencies in the published data [Li10,Barak07, Hay10].
- Construct varying bin-width histograms from previously released data[Machanavajjhala08], synthetic data[Xiao11], and from an equi-width histogram[Xu12].

Instead of adding noises to the frequency counts, can we publish the data directly?

Our Approach: main idea

- Sort the data; add noise directly to the data; and publish the noisy data.

D

↓ sorting

$$S(D) = \langle x_1 , x_2, x_3, \ldots , x_m \rangle$$

↓ add Laplace noise

$$S(D)' = \langle x_1 + n_1, x_2 + n_2, x_3 + n_3, \ldots , x_m + n_m \rangle$$

Our Approach: main idea

- Sort the data; add noise directly to the data; and publish the noisy data.

D

Will the published data too noisy?

↓

$S(D) = \langle x_1 , x_2, x_3, \ldots , x_m \rangle$

↓ add Laplace noise

$S(D)' = \langle x_1 + n_1, x_2 + n_2, x_3 + n_3, \ldots , x_m + n_m \rangle$

Our Approach: main idea

- Sort the data; add noise directly to the data; and publish the noisy data.

D

Will the published data too noisy?

↓

$S(D) = \langle x_1 , x_2, x_3, \ldots , x_m \rangle$

How to extend to higher dimension?

↓

$S(D)' = \langle x_1 + n_1, x_2 + n_2, x_3 + n_3, \ldots , x_m + n_m \rangle$

Observations & Techniques

1. Show that the sensitivity of “sorting” is not too large.

2. Exploit redundancy using Isotonic regression.

3. Grouping to tradeoff generalization errors with the level of Laplace noise.

4. Extension to higher dimension through location preservation mapping.

1. Sensitivity

For two neighbouring D and D' $\subset$ [0,1]

Sort(D)= $\langle x_1 \quad x_2 \quad x_3 \quad x_4 \quad \ldots \quad x_{m-2} \quad x_{m-1} \quad x_m \rangle$

Sort(D')= $\langle x_1 \quad x_3 \quad x_4 \quad \ldots \quad x_{m-2} \quad x_{m-1} \quad x_m \quad 1 \rangle$

1. Sensitivity

For two neighbouring D and D’ $\subset$ [0,1]

Sort(D)= $\langle\ x_1\quad x_2\quad x_3\quad x_4\quad \ldots\quad x_{m-2}\quad x_{m-1}\quad x_m\ \rangle$

Sort(D’)= $\langle\ x_1\quad x_3\quad x_4\quad \ldots\quad x_{m-2}\quad x_{m-1}\quad x_m\quad 1\ \rangle$

$|$ sort (D) – sort (D’) $|_1$

1. Sensitivity

For two neighbouring D and D' $\subset [0,1]$

Sort(D)= $\langle\ x_1 \quad x_2 \quad x_3 \quad x_4 \quad \ldots \quad x_{m-2} \quad x_{m-1} \quad x_m\ \rangle$

Sort(D')= $\langle\ x_1 \quad x_3 \quad x_4 \quad \ldots \quad x_{m-2} \quad x_{m-1} \quad x_m \quad 1\ \rangle$

$$| \text{sort}(D) - \text{sort}(D') |_1 \leq 1$$

2. Isotonic regression

Note that the sorted data are constrained: the elements are increasing.

Isotonic regression: Given a sequence

$Y = \langle y_1, y_2, y_3, \ldots, y_m \rangle$

find an non-decreasing sequence

$X = \langle x_1, x_2, x_3, \ldots, x_m \rangle$

minimizing the distance of X from Y.

3. Grouping

Group consecutive elements and publish its noisy sum.

Sort(D)= $\langle\ x_1 \quad x_2 \quad x_3 \quad x_4 \quad \ldots \quad x_{m-2} \quad x_{m-1} \quad x_m\ \rangle$

$x_1+x_2+x_3$ $\quad$ $x_4+x_5+x_6$ $\quad \ldots \quad$ $x_{m-2}+x_{m-1}+x_m$

3. Grouping

Group consecutive elements and publish its noisy sum.

$$\text{Sort}(D) = \langle\ x_1 \quad x_2 \quad x_3 \quad x_4 \quad \ldots \quad x_{m-2} \quad x_{m-1} \quad x_m\ \rangle$$

$$x_1+x_2+x_3 \qquad x_4+x_5+x_6 \qquad \ldots \qquad x_{m-2}+x_{m-1}+x_m$$

Grouping does not affect sensitivity.

3. Grouping

Group consecutive elements and publish its noisy sum.

$$\text{Sort}(D) = \langle\, x_1 \quad x_2 \quad x_3 \quad x_4 \quad \ldots \quad x_{m-2} \quad x_{m-1} \quad x_m \,\rangle$$

$x_1+x_2+x_3$ + noise

$x_4+x_5+x_6$ + noise

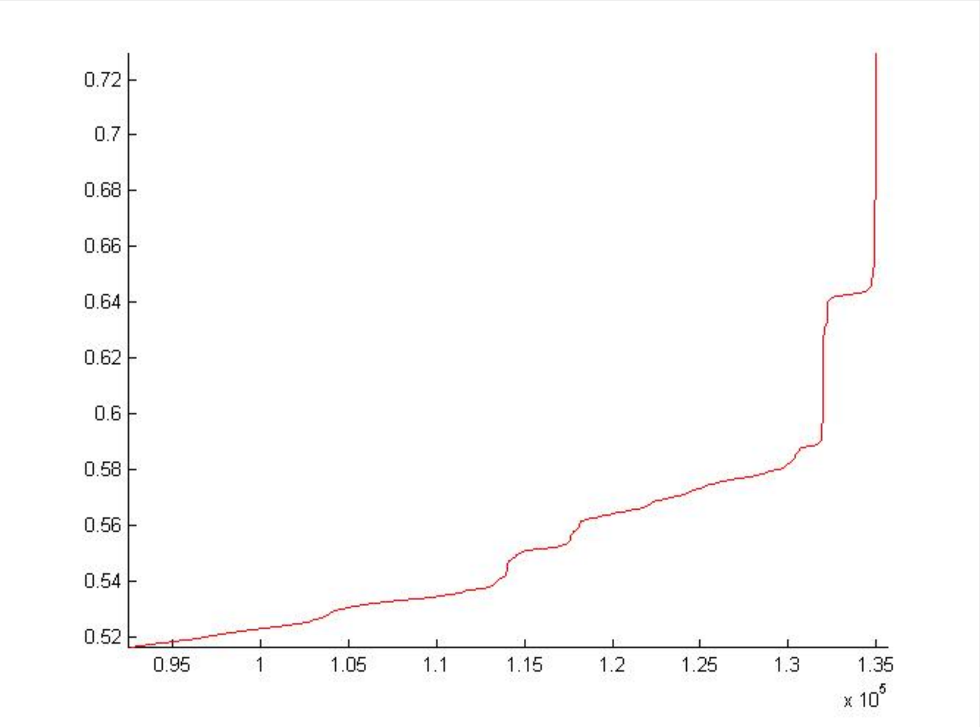
…

$x_{m-2}+x_{m-1}+x_m$ + noise

Grouping does not affect sensitivity.

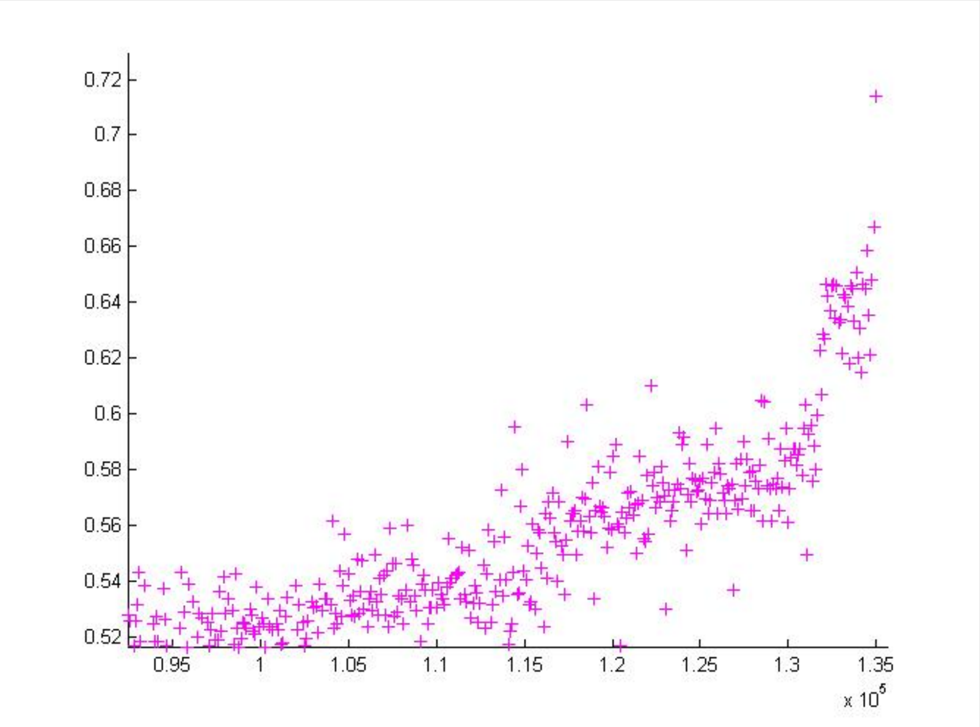
Illustration

The Grouped Sorted data



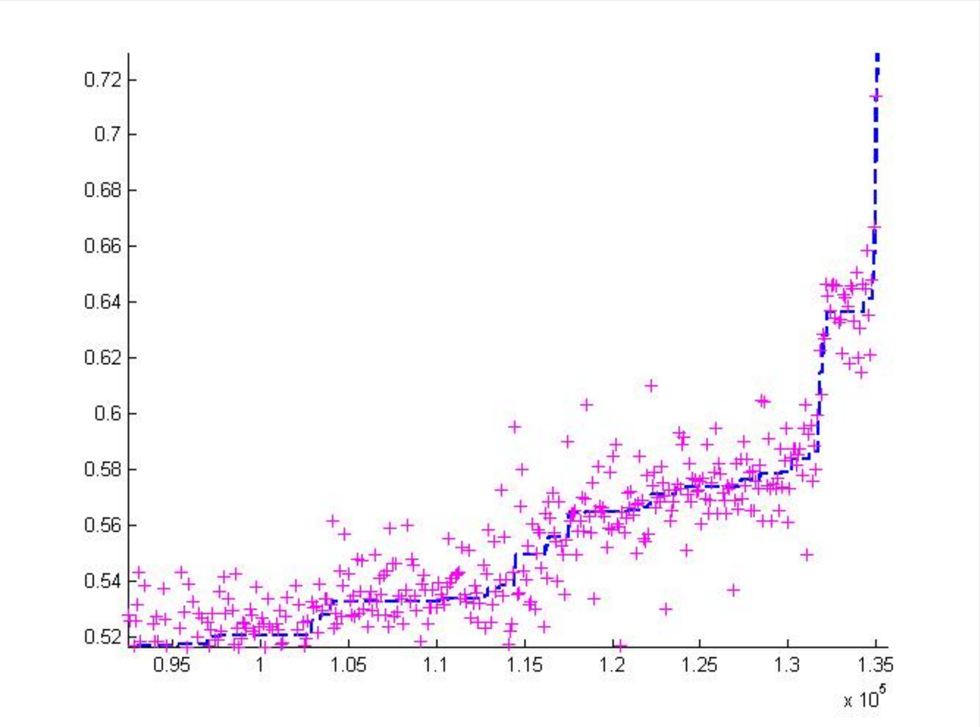
Illustration

With Laplace Noise



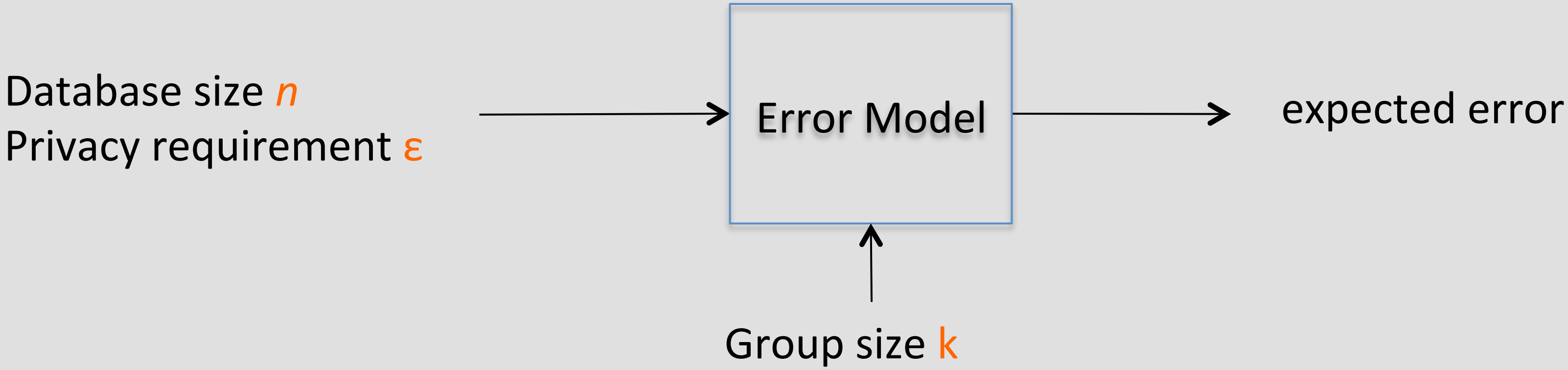
Illustration

Isotonic regression.

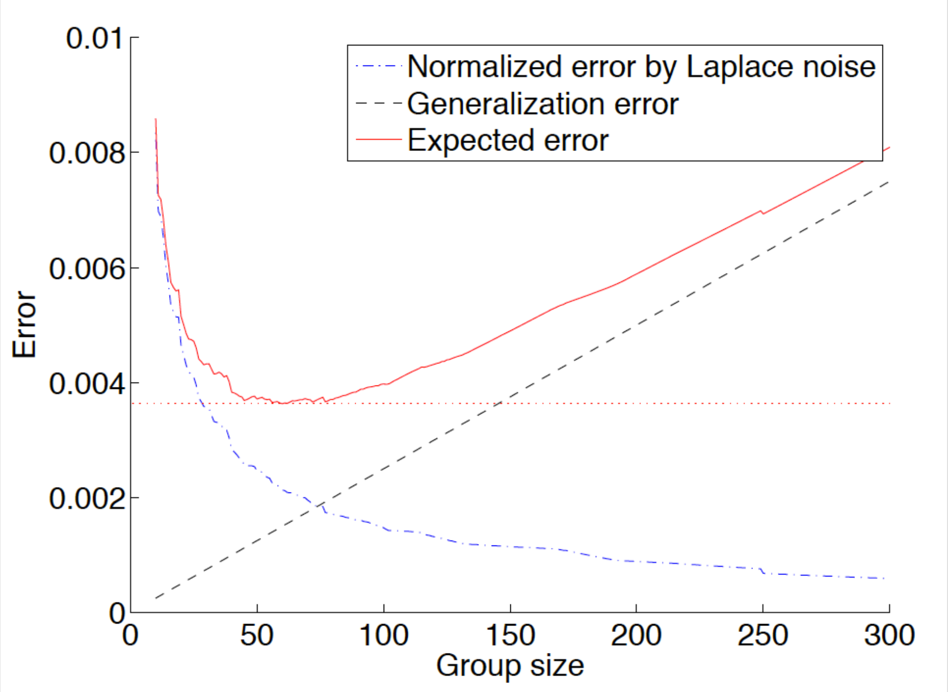


Grouping: what should be the appropriate group size?

We give a model to estimate the expected error based on the (1) group size *k*, (2) size of dataset *n* and (3) privacy requirement ε.



From the model, we can estimate the optimal group size k, given *n* and ε .

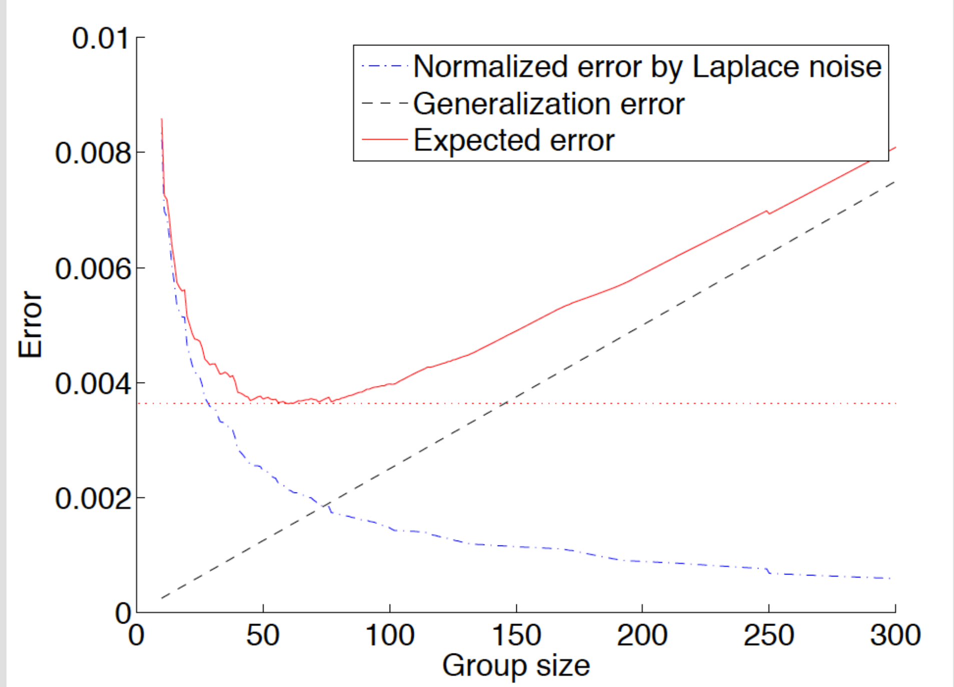


Expected_error (ε, k, n) ≈

Generalization_error (n, k)

+

Laplace_noise (ε, n, k)

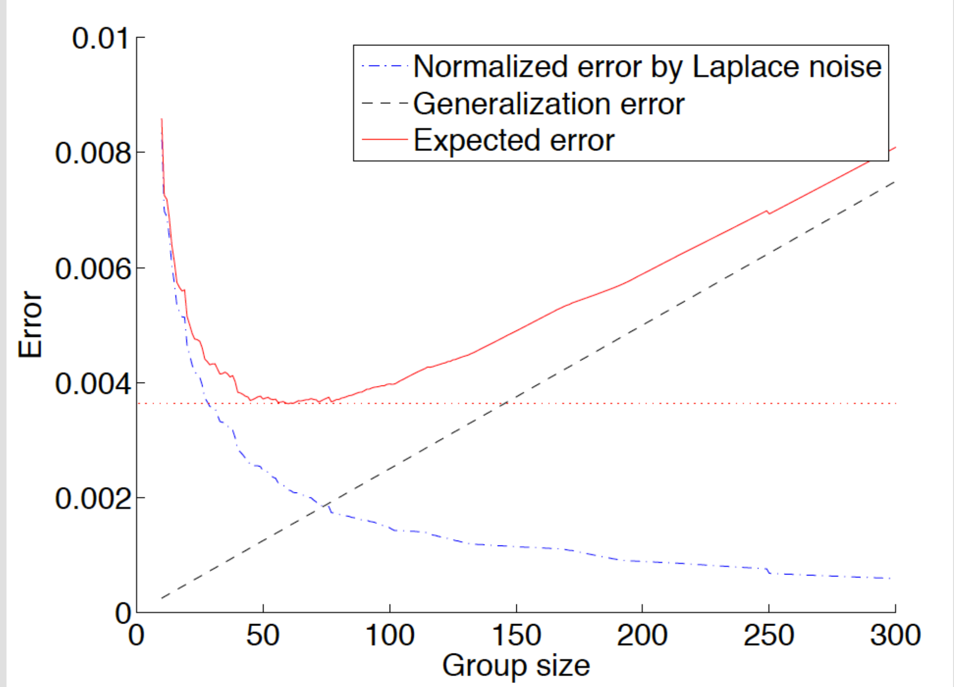


$$\text{Expected\_error}(\varepsilon, k, n) \approx$$

$$\text{Generalization\_error}(n, k)$$

$$+$$

$$k^{-1}\ \text{Laplace\_noise\_without\_grouping}(\varepsilon, n\,k^{-1})$$



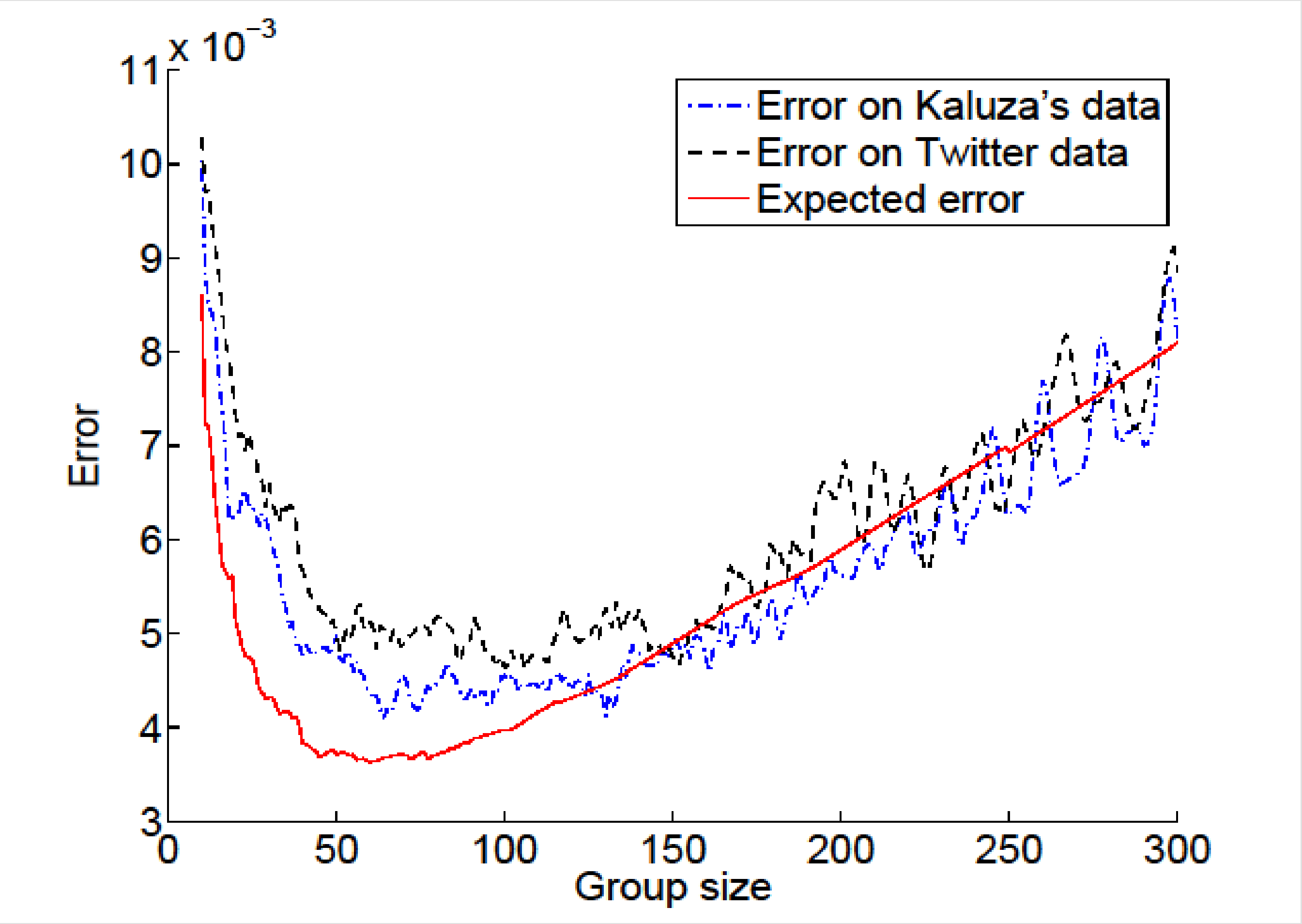
Expected_error (ε, k, n) ≈

Generalization_error (n, k)

+

k^{-1} Laplace_noise_without_grouping (ε, n k^{-1})

Accuracy of Error Model



Kaluza's data: [Kaluza10]

Twitter data:[Twitterdata10]

4. Extension to Higher Dimension

- Consider location preserving mapping

  $T: [0,1]\times[0,1] \rightarrow [0,1]$

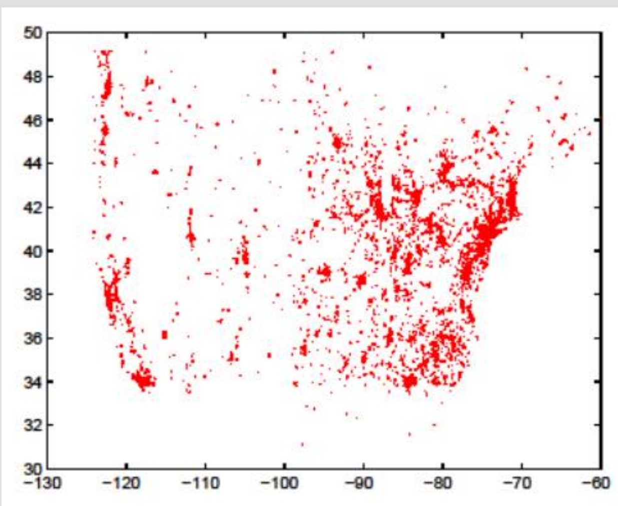
  s.t.,

  if T(x) and T(y) are "close-by" in [0,1]

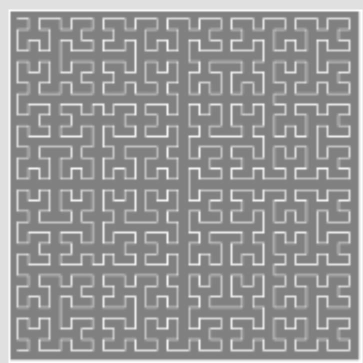
  then x, y are "close-by" in [0,1]×[0,1]

Extension to Higher Dimension

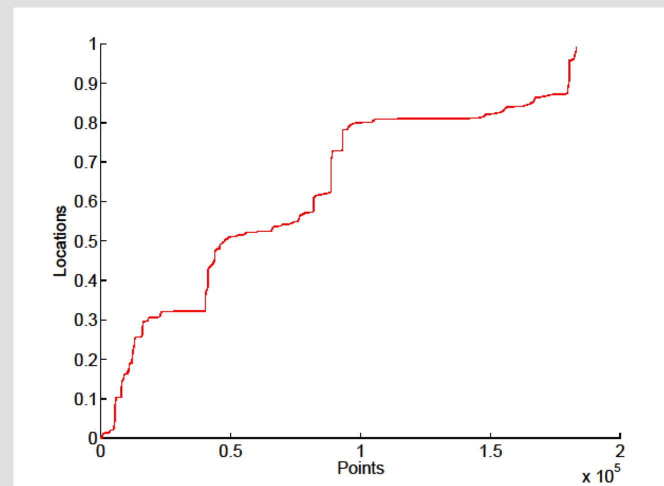
- Example of such mapping: Hilbert space filling curve.



2D data points



Location preserving mapping



Sorted 1D data points

Putting all together: Proposed mechanisms

Given the dataset D, privacy requirement ε, the publisher performs:

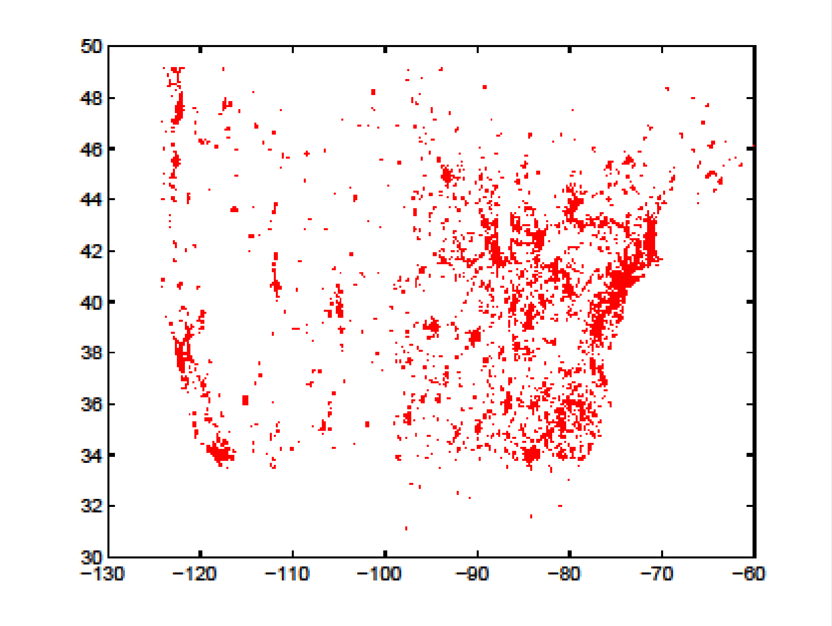
1. Determines the group size *k* from n=|D|, and ε .
2. Maps D to [0,1]. Let the mapped points be T(D).
3. Sorts T(D).
4. Groups k consecutive elements.
5. Adds noise to the sum in each group. Publishes the noisy sums.

Given the published data, a user performs:

1. Isotonic regression.
2. Inverse of the location preserving mapping.
3. Subsequent operations, like query, visualization, & data mining.

Evaluation: Datasets

- Profile of Twitter users.[Twitterdata10]

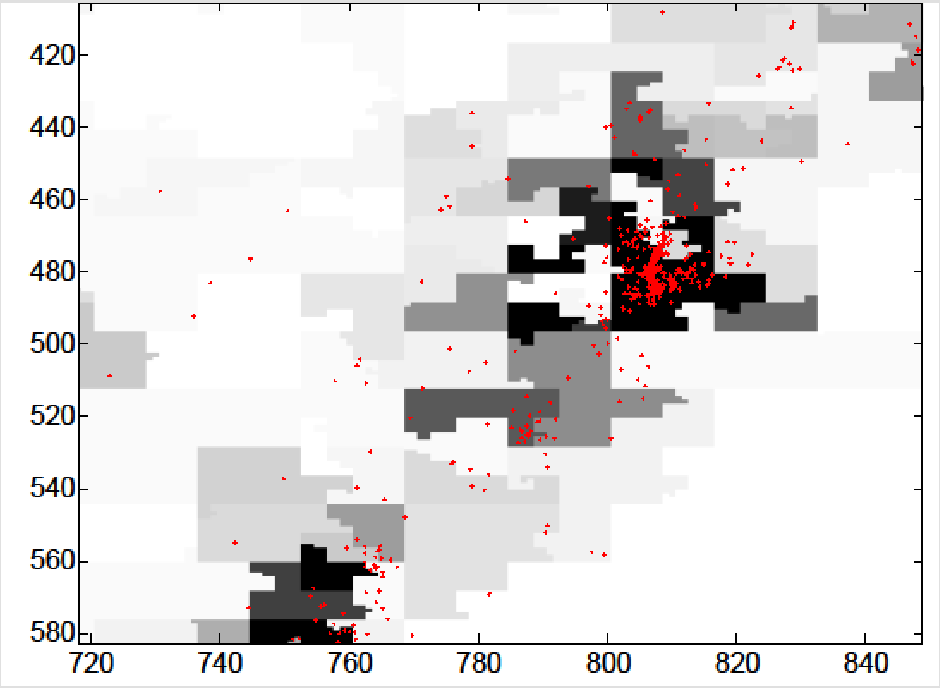


Locations of 180,000 profiles in North America.

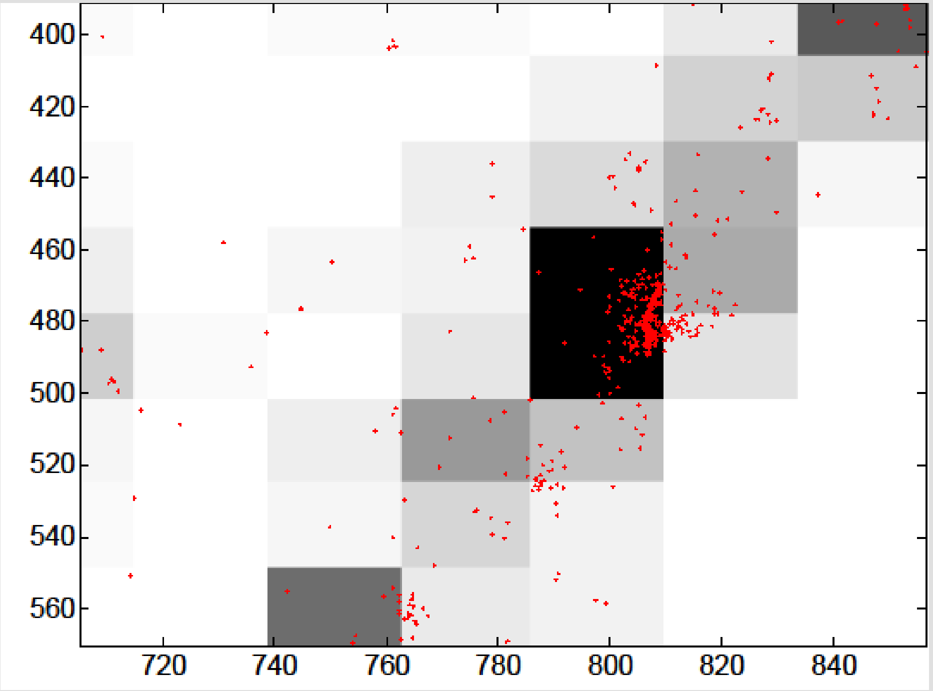
- The distance of the locations to New York City is taken as the 1D data.

Adaptive Resolution

- A visualization of our method and equi-width histogram

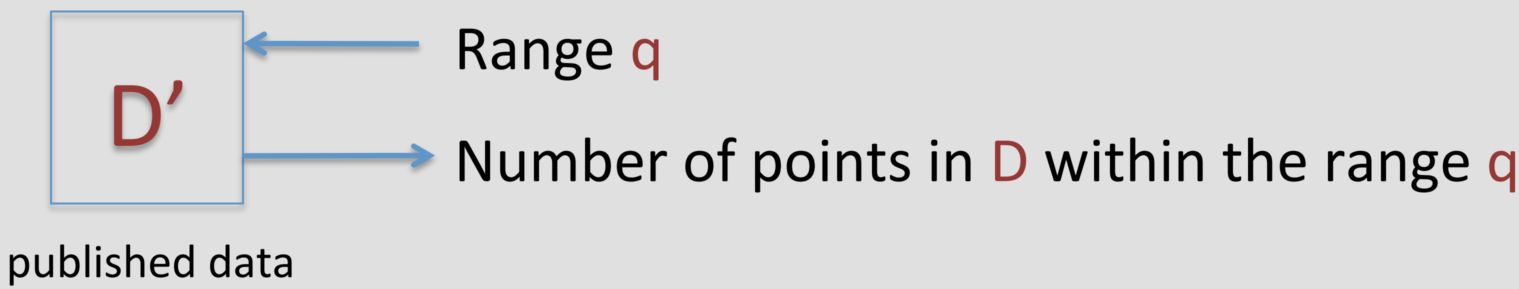


Our method



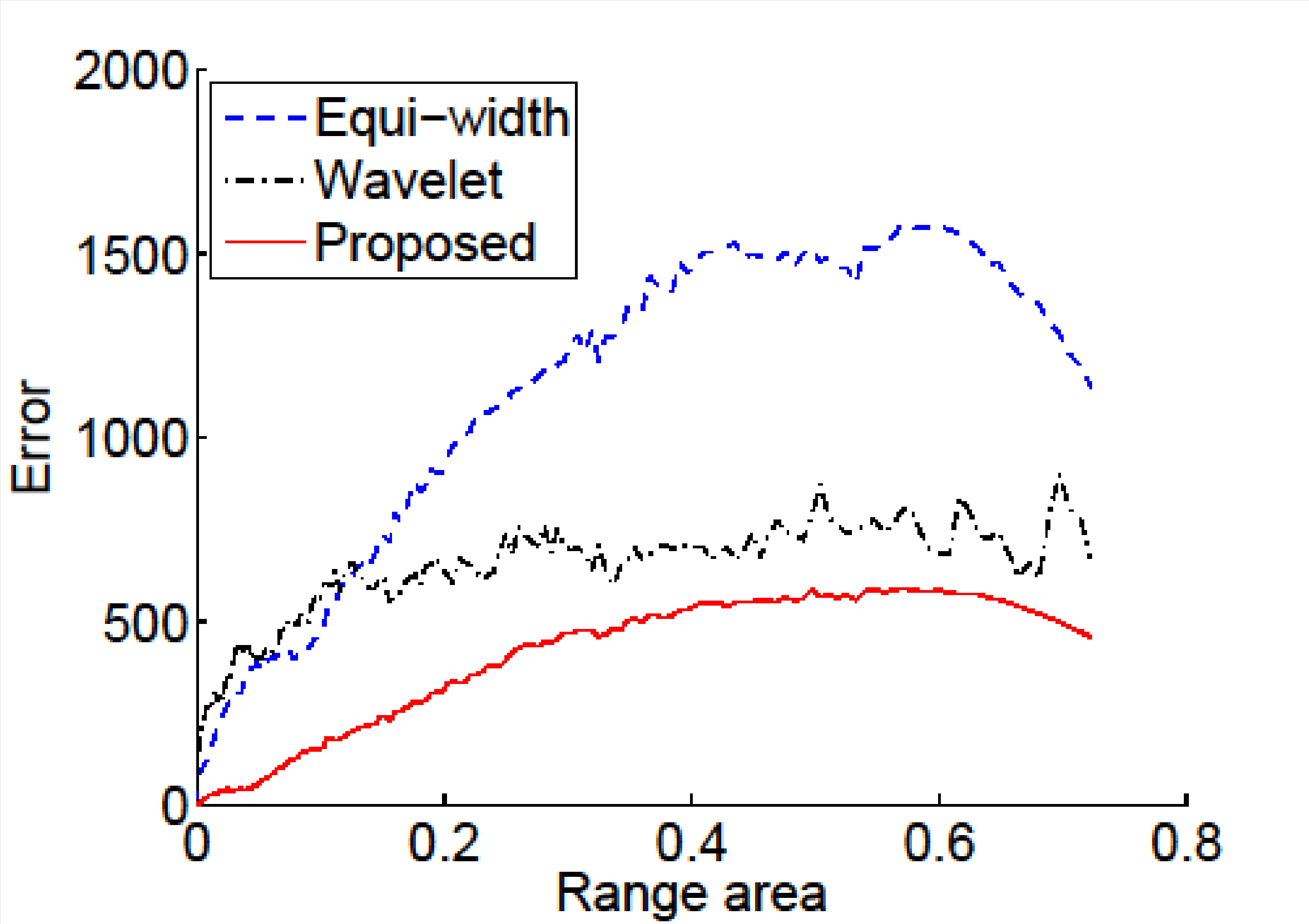
Equi-width histogram

Evaluation: Range Query

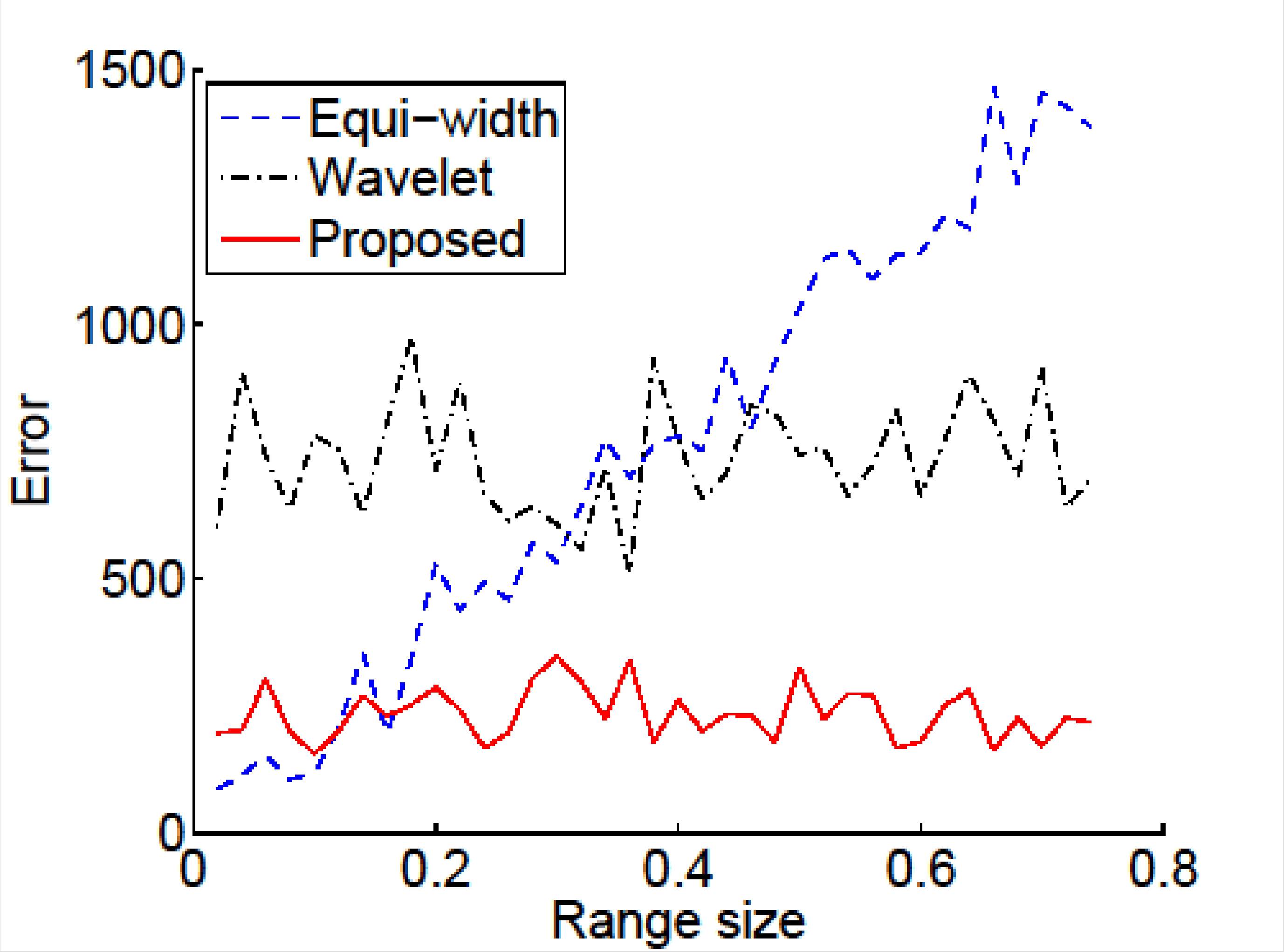


- We repeat the experiment 1,000 times for each size of the range q. We compare our algorithm with equi-width histogram and wavelet-based method [Xiao10].

Range Query: 2D domain



Range Query: 1D domain



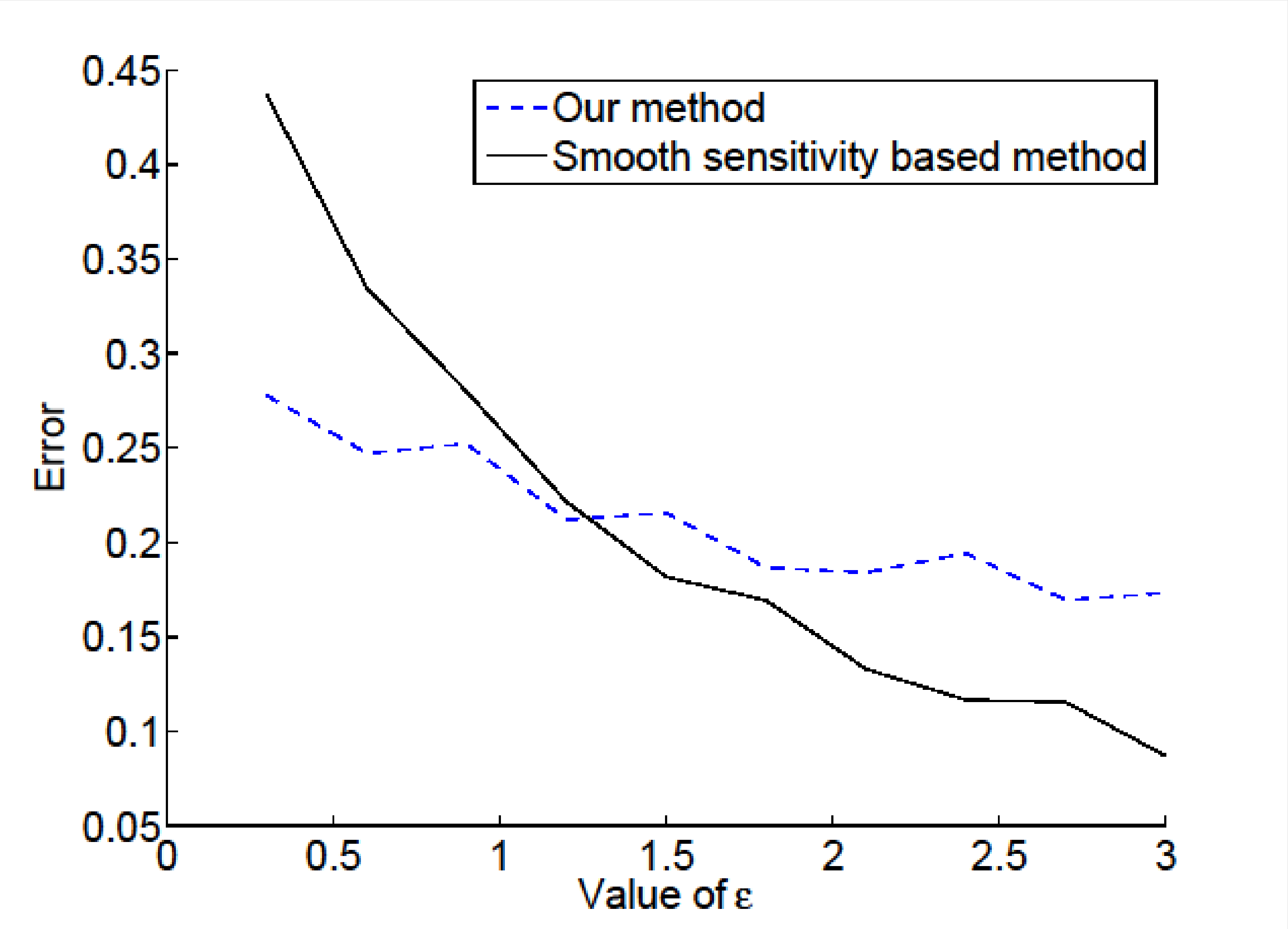
Evaluation: Median-Finding

$D = d_1 \dots d_m \dots d_{2m-1}$ → d_m published median

Original sorted data

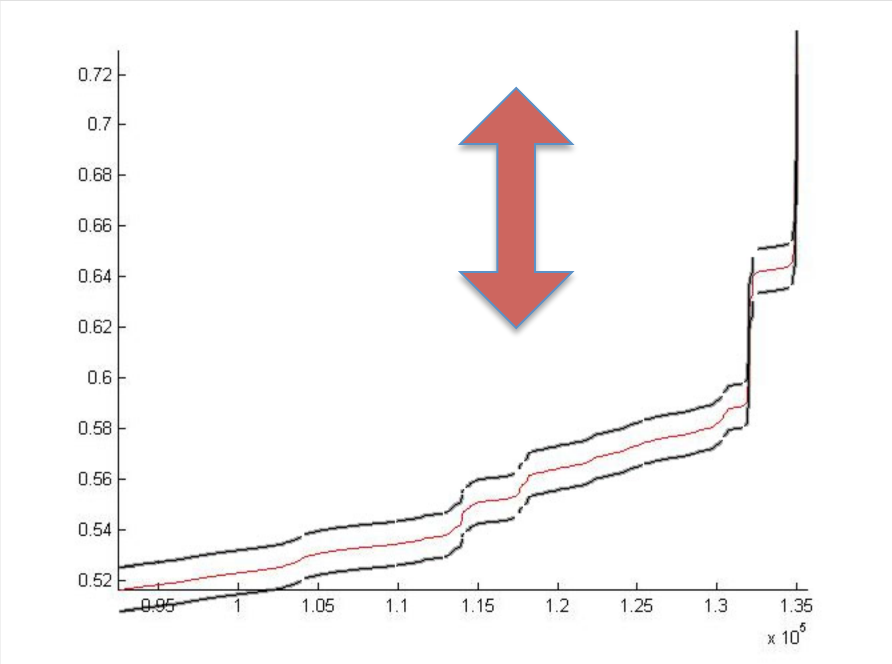
- We compare our algorithm with the smooth-sensitivity approach [Nissim07].

Evaluation: Median-finding

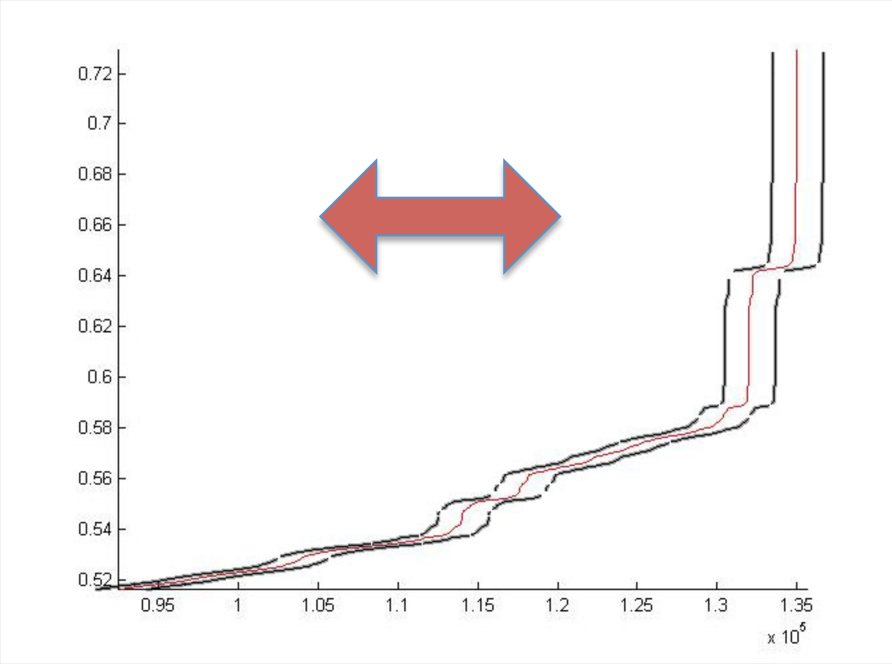


Discussion: Complementary

- Alternative ``direction'' of the Laplace noise



Our method



Equi-width histogram

Conclusion

- We proposed an approach that publishes the data directly.
  - Simple.
  - The main parameter (group size) can be determined without the dataset D. In contrast, optimal parameters of many existing mechanisms heavily rely on the dataset.
  - Leads to adaptive histograms. Achieve high utility.
  - Complementary to the frequency-counts methods and potentially can be combined for higher utility.

- We proposed using location preservation mapping for extension to low-dimensional data (for e.g. 2D and 3D).

Reference

- [Xiao10]: X. Xiao, G. Wang, and J. Gehrke. *Differential privacy via wavelet transforms*. IEEE Transactions on Knowledge and Data Engineering, page 1200, 2010.
- [Hay10]: M. Hay, V. Rastogi, G. Miklau, and D. Suciu. *Boosting the accuracy of dierffentially private histograms through consistency*. VLDB Endowment, page 1021, 2010.
- [Chan11]: T.H.H. Chan, E. Shi,and D. Song. *Private and continual release of statistics*. ACM Transactions on Information and System Security, page 26, 2011.
- [Li10]: C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. *Optimizing linear counting queries under differential privacy*. ACM Symposium on Principles of Database Systems of Data, page 123, 2010.
- [Barak07]: B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. *Privacy, accuracy, and consistency too: a holistic solution to contingency table release*. symposium on principles of database systems, page 273, 2007.
- [Machanavajjhala08]: A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. *Privacy: Theory meets practice on the map*. In International Conference on Data Engineering, page 277, 2008.
- [Xiao11]: Y. Xiao, L. Xiong, and C. Yuan. *Dierentially private data release through multidimensional partitioning*. Secure Data Management, page 150, 2011.
- [Xu12]: J. Xu, Z. Zhang, X. Xiao, Y. Yang, and G. Yu. *Differentially Private Histogram Publication*. IEEE International Conference on Data Engineering (ICDE), page 32, 2012.
- [Dwork06]: C. Dwork. *Differential privacy.* Automata, languages and programming, page 1, 2006.
- [Nissim07]: K. Nissim, S. Raskhodnikova, and A. Smith. *Smooth sensitivity and sampling in private data analysis*. ACM Symposium on Theory of Computing, page 75, 2007.
- [Kaluza10]: B. Kaluza, V. Mirchevska, E. Dovgan, M. Lustrek, and M. Gams. *An agent-based approach to care in independent living*. Ambient Intelligence, page 177, 2010.
- [Twitterdata10]: First 180k profiles from http://www.infochimps.com/datasets/twitter-census-twitter-users-by-location .