# Problems with Differential Privacy

Presenter: Guoming Wang

3 March, 2015

Nanyang Technological University

# Materials

1. Larry Wasserman: “A Statistical View of Differential privacy”. Carnegie Mellon University.

## Challenge -- Solution: Compress the sequence

### *Reduce effective N by compressing the sequence*

***Discrete Fourier Transform (DFT):***

$q_1,\ldots,q_N \xrightarrow{\text{DFT}} f_1,\ldots,f_N \xrightarrow{\text{Inverse DFT}} q_1,\ldots,q_N$

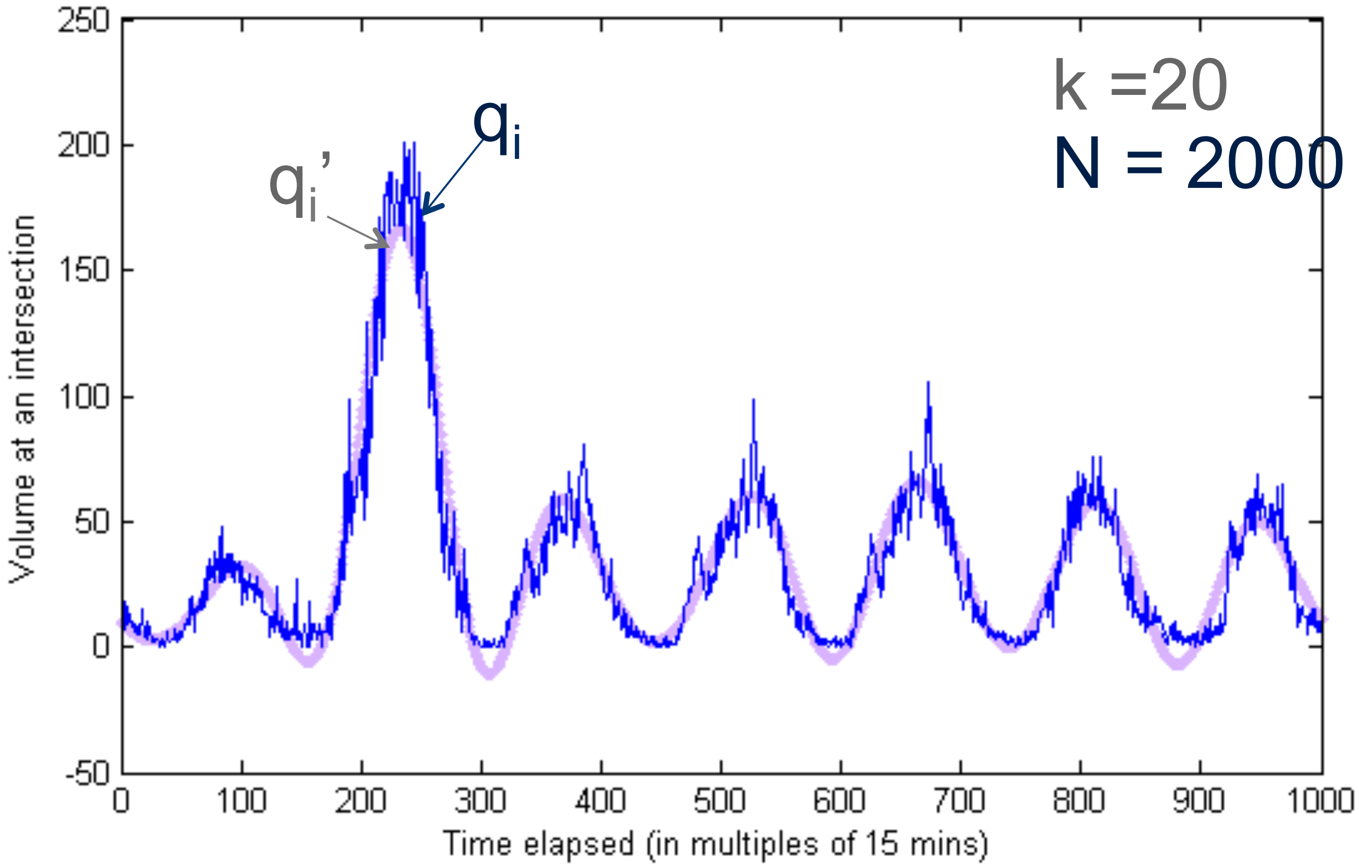
***DFT-based Compression (NOT private):***

$q_1,\ldots,q_N \xrightarrow{\text{DFT}} f_1,\ldots,f_k,0,\ldots,0 \xrightarrow{\text{Inverse DFT}} q'_1,\ldots,q'_N$

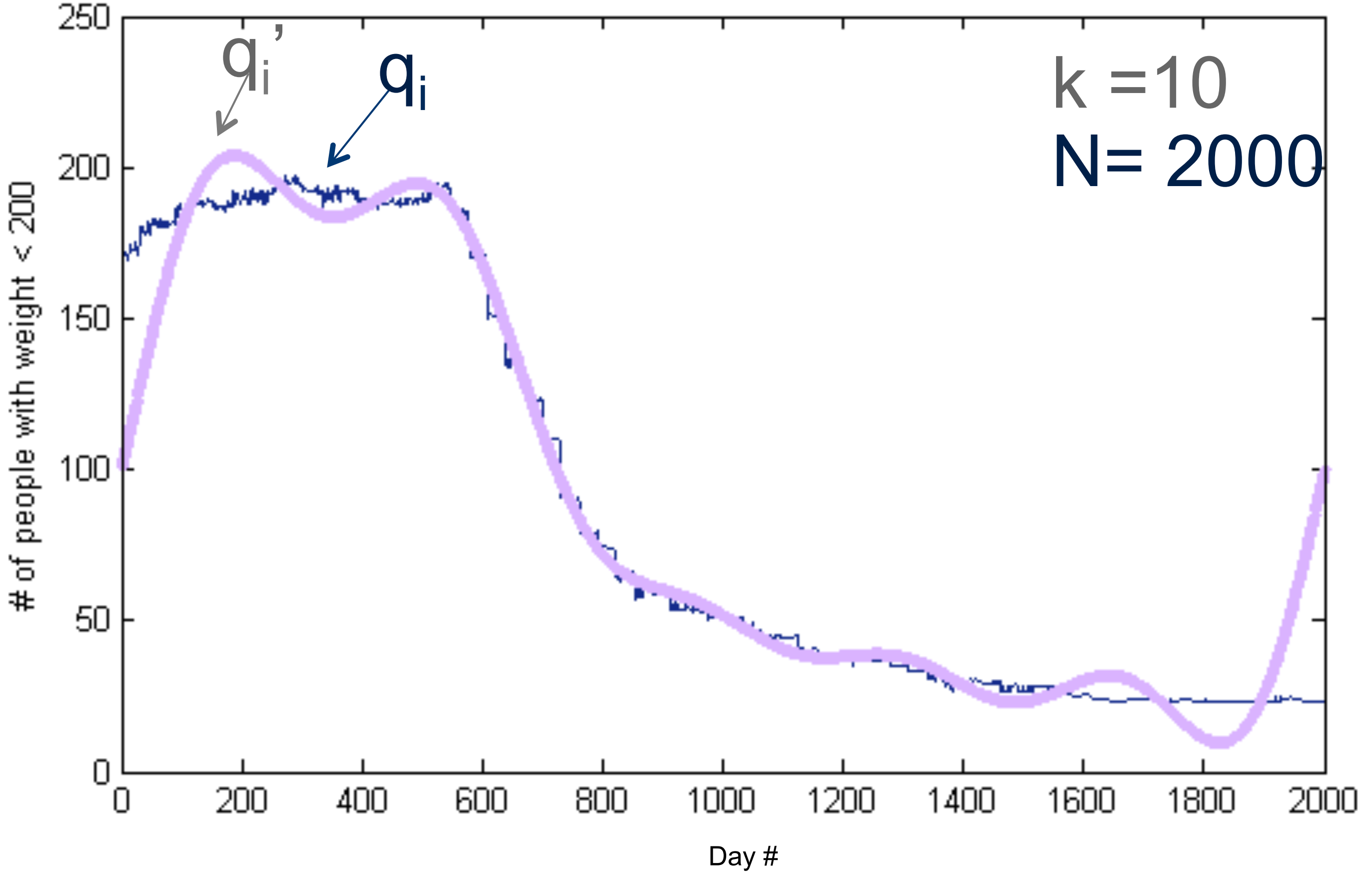
$q'_i$ has some error compared to $q_i$

- Error is small if $q_i$ has periodic nature
- k/N is the compression ratio

## DFT-based Compression - Examples



## DFT-based Compression - Examples



# Outline

- **Different view of CS and Statistics**
- **Problems of Differential Privacy**

## CS vs Statistics

What they do:

# Statisticians: mostly applied statisticians working on real problems.

Want methods that worked on real, complex, data sets.

# CS People: mainly theoreticians doing very interesting theory.

Want precise definitions of privacy and theorems guaranteeing that privacy held

## CS vs Statistics

What they do:

# Statisticians: mostly applied statisticians working on real problems.

* Want methods that worked on real, complex, data sets.

* give me data, Then I can: draw plots, fit models, test fit, estimate parameters, make predictions, …

# CS People: mainly theoreticians doing very interesting theory.

* Want precise definitions of privacy and theorems guaranteeing that privacy held.

* receive a query, return a private answer.

## Statistical Concepts

- Data D = (X1, …, Xn) where X1, …Xn ~ P.
- Open view the database as a sample from a population. The goal is not just to summarize the database; they want to infer (learn) about the population.
- Formally, the goal is to infer P or some functions of P (means, correlations, etc.) or predict a new observation.

1. Data (X1, Y1), … , (Xn, Yn).
2. Observe new Xn+1. Predict Yn+1.
3. If Y blongs to R this is regression. If Y is discrete this is classification.

## Problems with Differential Privacy

- Differential Privacy is a precise and strong guarantee.
- But there are two problems:
  - Recall that X = set of possible databases. X is ambiguous.
  - The notion of neighboring databases can be ambiguous.
- In many real problems, it simply is not clear what x is.
- Need to know x to even implement differential privacy.

## Problems with Differential Privacy

- Second, it is too strong.
- Consider a high dimensional contingency table. The counts are very sparse. There are many zeroes.
- The sample size n is much smaller than the number of cells.
- Creating a synthetic database subject to differential privacy leads to a very noisy database. (Mostly noise.)

## Conclusion

- Differential Privacy is a precise, mathematical guarantee.
- Useful theoretically but makes it somewhat impractical.
- Mostly ignored statistics.

# Thank you – Enjoy the rest of your night