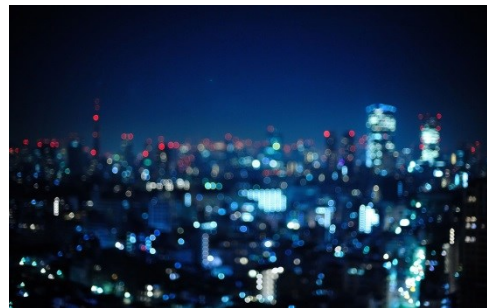


The confounding problem of private data release

Graham Cormode

g.cormode@warwick.ac.uk



Big data, big problem?

- ◆ The **big data meme** has taken root
 - Organizations jumped on the bandwagon
 - Funding agencies have given out grants
- ◆ But the data comes from **individuals**
 - Individuals want privacy for their data
 - How can scientists work on sensitive data?
- ◆ The **easy answer**: **anonymize it** and release
- ◆ The **problem**: we don't know how to do this



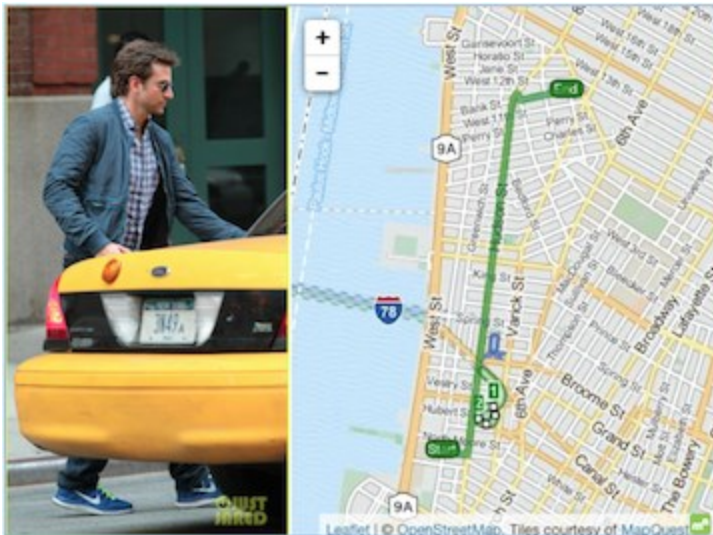
A recent data release example

- ◆ NYC taxi and limousine commission released 2013 trip data
 - Contains start point, end point, timestamps, taxi id, fare, tip amount
 - 173 million trips “anonymized” to remove identifying information
- ◆ **Problem:** the anonymization was easily reversed
 - Anonymization was a simple hash of the identifiers
 - Small space of ids, easy to brute-force dictionary attack
- ◆ But so what?
 - Taxi rides aren't sensitive?



Almost anything can be sensitive

- ◆ Can link people to taxis and find out where they went
 - E.g. paparazzi pictures of celebrities



Bradley Cooper (actor)



Jessica Alba (actor)

Finding sensitive activities



- ◆ Find trips starting at remote, “sensitive” locations
 - E.g. Larry Flynt’s Hustler Club [an “adult entertainment venue”]
- ◆ Can find where the venue’s customers live with high accuracy
 - “Examining one of the clusters revealed that only one of the 5 likely drop-off addresses was inhabited; a search for that address revealed its resident’s name.
In addition, by examining other drop-offs at this address, I found that this gentleman also frequented such establishments as “Rick’s Cabaret” and “Flashdancers”.
Using websites like Spokeo and Facebook, I was also able to find out his property value, ethnicity, relationship status, court records and even a profile picture!”

◆ Oops

We've heard this story before...



We need to solve this data release problem...



Crypto is not the (whole) solution

- ◆ **Security is binary**: allow access to data **iff** you have the key
 - Encryption is robust, reliable and widely deployed
- ◆ **Private data release comes in many shades**:
reveal some information, disallow unintended uses
 - Hard to control what may be inferred
 - Possible to combine with other data sources to breach privacy
 - Privacy technology is still maturing
- ◆ **Goals for data release**:
 - Enable appropriate use of data while protecting data subjects
 - Keep chairman and CTO off front page of newspapers
 - Simplify the process as much as possible: 1-click privacy?



PAST: PRIVACY AND THE DB COMMUNITY

Aspects of Privacy

- ◆ **First-person privacy**: Who can see what about me?
 - **Example**: Who can see my holiday photos on a social network?
 - **Failure**: “Sacked for complaining about boss on Facebook!”
 - **Controls**: User sets up rules/groups for other (authenticated) users
- ◆ **Second-person privacy**: Who can share your data with others?
 - **Example**: Does a search engine share your queries with advertisers?
 - **Failure**: MySpace leaks user ids to 3rd party advertisers
 - **Controls**: Policy, regulations, scrutiny, “Do Not Track”
- ◆ **Third-person (plural) privacy**: Can you be found in the crowd?
 - **Example**: Can trace someone’s movements in a mobility dataset?
 - **Failure**: AOL releases search logs that allow users to be identified
 - **Controls**: Access controls and anonymization technology

Example Business Payment Dataset

Name	Address	DOB	Sex	Status
Fred Bloggs	123 Elm St, 53715	1/21/76	M	Unpaid
Jane Doe	99 MLK Blvd, 53715	4/13/86	F	Unpaid
Joe Blow	2345 Euclid Ave, 53703	2/28/76	M	Often late
John Q. Public	29 Oak Ln, 53703	1/21/76	M	Sometimes late
Chen Xiaoming	88 Main St, 53706	4/13/86	F	Pays on time
Wanjiku	1 Ace Rd, 53706	2/28/76	F	Pays on time

- ◆ **Identifiers**—uniquely identify, e.g. Social Security Number (SSN)
- ◆ **Quasi-Identifiers (QI)**—such as DOB, Sex, ZIP Code
- ◆ **Sensitive attributes (SA)**—the associations we want to hide

Deidentification

Address	DOB	Sex	Status
123 Elm St, 53715	1/21/76	M	Unpaid
99 MLK Blvd, 53715	4/13/86	F	Unpaid
2345 Euclid Ave, 53703	2/28/76	M	Often late
29 Oak Ln, 53703	1/21/76	M	Sometimes late
88 Main St, 53706	4/13/86	F	Pays on time
1 Acer Rd, 53706	2/28/76	F	Pays on time

Anonymized?

Post Code	DOB	Sex	Status
53715	1/21/76	M	Unpaid
53715	4/13/86	F	Unpaid
53703	2/28/76	M	Often late
53703	1/21/76	M	Sometimes late
53706	4/13/86	F	Pays on time
53706	2/28/76	F	Pays on time

Generalization and k-anonymity

Post Code	DOB	Sex	Status
537**	1/21/76	M	Unpaid
537**	4/13/86	F	Unpaid
537**	2/28/76	*	Often late
537**	1/21/76	M	Sometimes late
537**	4/13/86	F	Pays on time
537**	2/28/76	*	Pays on time

Definitions in the literature





PRESENT: SOME STEPS TOWARDS PRIVACY

Differential Privacy (Dwork et al 06)

A randomized algorithm K satisfies ϵ -differential privacy if:

Given two data sets that differ by one individual, D and D' , and any property S :

$$\Pr[K(D) \in S] \leq e^\epsilon \Pr[K(D') \in S]$$

- Can achieve differential privacy for counts by adding a random noise value
- Uncertainty due to noise “hides” whether someone is present in the data

Achieving ϵ -Differential Privacy

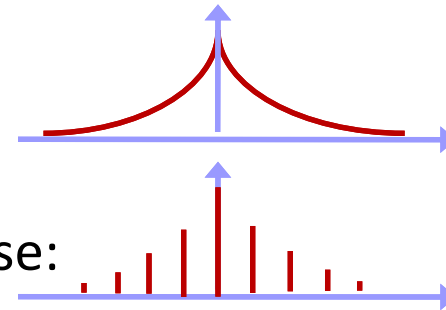
(Global) Sensitivity of publishing:

$$s = \max_{x, x'} |F(x) - F(x')|, x, x' \text{ differ by 1 individual}$$

E.g., count individuals satisfying property P : one individual changing info affects answer by at most 1; hence $s = 1$

For every value that is output:

- Add Laplacian noise, $\text{Lap}(\epsilon/s)$:
- Or Geometric noise for discrete case:



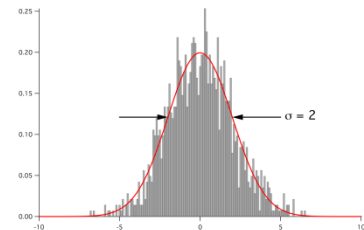
Simple rules for composition of differentially private outputs:

Given output O_1 that is ϵ_1 private and O_2 that is ϵ_2 private

- (Sequential composition) If inputs overlap, result is $\epsilon_1 + \epsilon_2$ private
- (Parallel composition) If inputs disjoint, result is $\max(\epsilon_1, \epsilon_2)$ private

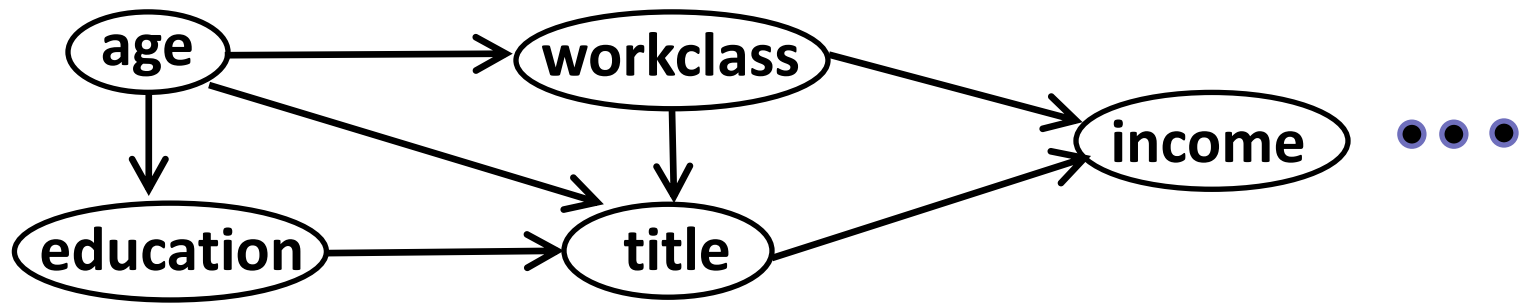
Differential privacy for data release

- ◆ Differential privacy is an attractive model for data release
 - Achieve a fairly robust statistical guarantee over outputs
- ◆ **Problem:** how to apply to data release where $f(x) = x$?
 - Trying to use global sensitivity does not work well
- ◆ **General recipe:** find a model for the data
 - Choose and release the model parameters under DP
- ◆ A new tradeoff in picking suitable models
 - Must be robust to privacy noise, as well as fit the data
 - Each parameter should depend only weakly on any input item
 - Need different models for different types of data
- ◆ Next 3 biased examples of recent work following this outline



Example 1: PrivBayes [SIGMOD 14]

- ◆ Directly materializing relational data: low signal, high noise
- ◆ Use a **Bayesian network** to approximate the full-dimensional distribution by lower-dimensional ones:



$$\begin{aligned} \Pr[H] \approx & \Pr[\text{age}] \cdot \Pr[\text{education}|\text{age}] \cdot \Pr[\text{workclass}|\text{age}] \cdot \\ & \Pr[\text{title}|\text{age,education,workclass}] \cdot \Pr[\text{income}|\text{workclass,title}] \cdot \\ & \Pr[\text{marital status}|\text{age,income}] \dots \end{aligned}$$

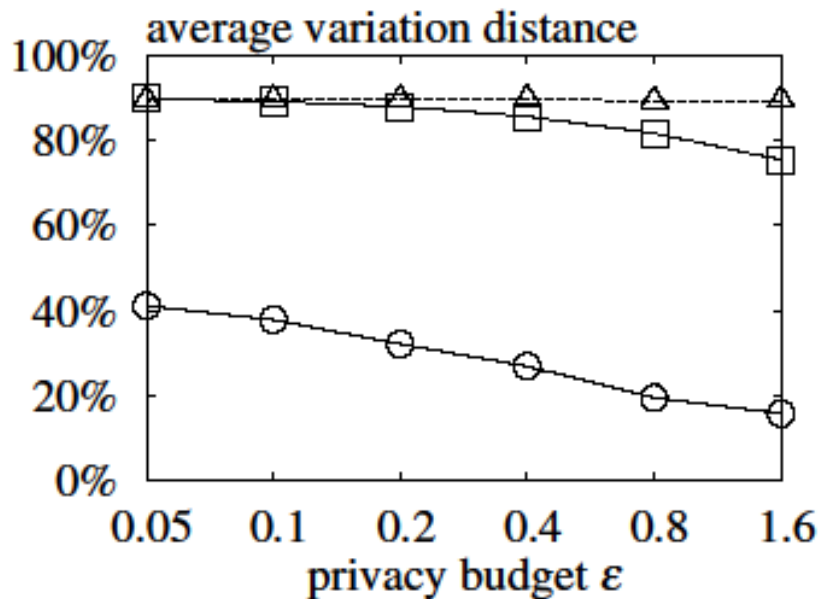
low-dimensional distributions: **high signal-to-noise**

PrivBayes (SIGMOD14)

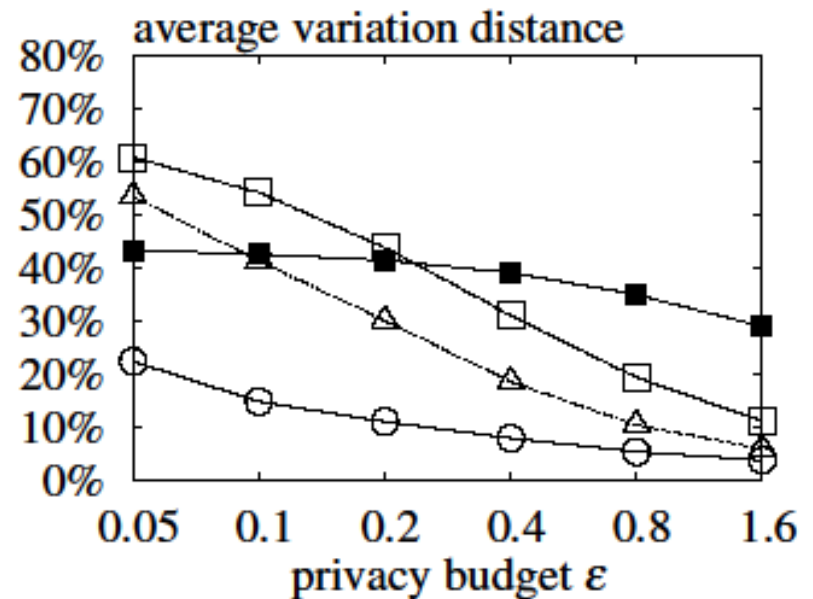
- ◆ **STEP 1:** Choose a suitable Bayesian Network BN
 - in a differentially private way
 - sample (via exponential mechanism) edges in the network
- ◆ **STEP 2:** Compute distributions implied by edges of BN
 - straightforward to do under differential privacy (Laplace)
- ◆ **STEP 3:** Generate synthetic data by sampling from the BN
 - post-processing: no privacy issues
- ◆ Evaluate utility of synthetic data for variety of different tasks
 - performs well for multiple tasks (classification, regression)

Experiments: Counting Queries

○ *PrivBayes* □ *Laplace* △ *Fourier* ■ *Histogram*



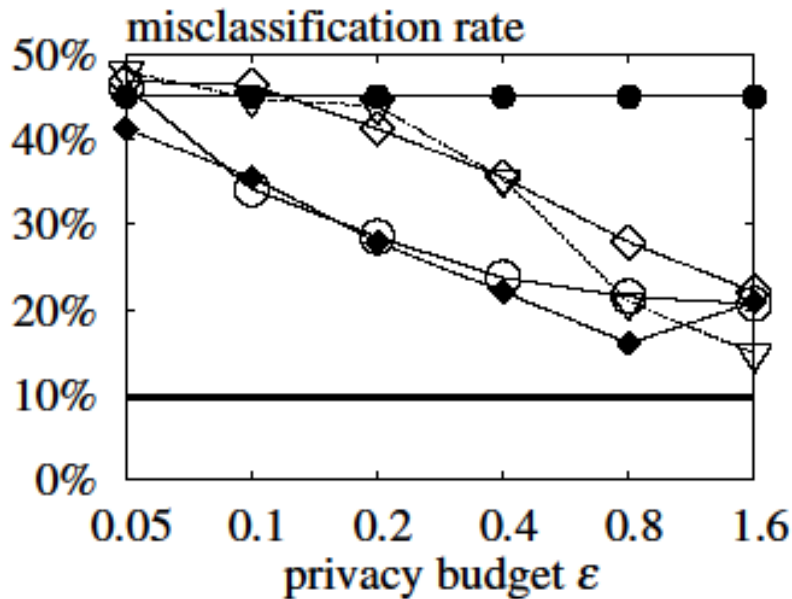
Adult dataset



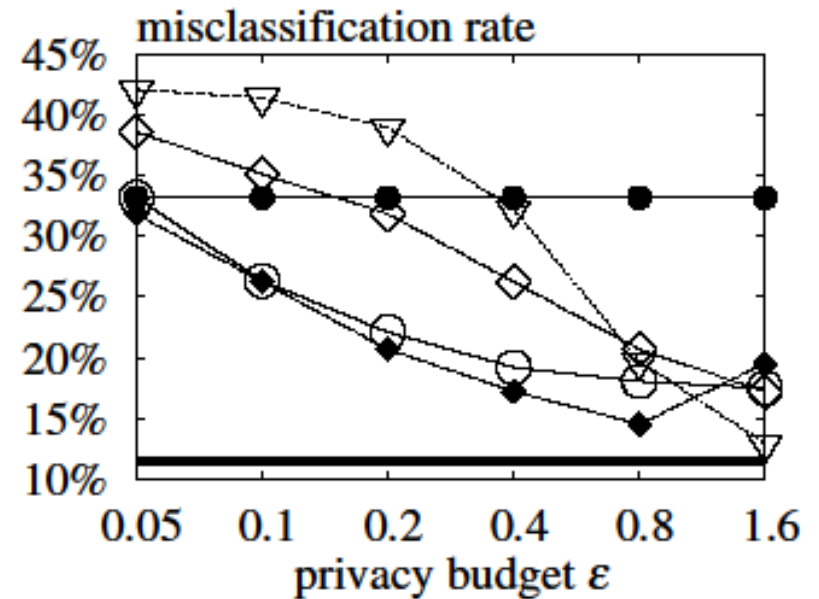
NLTCs dataset

Query load = Compute all 3-way marginals

Experiments: Classification



Y = education: post-secondary degree?

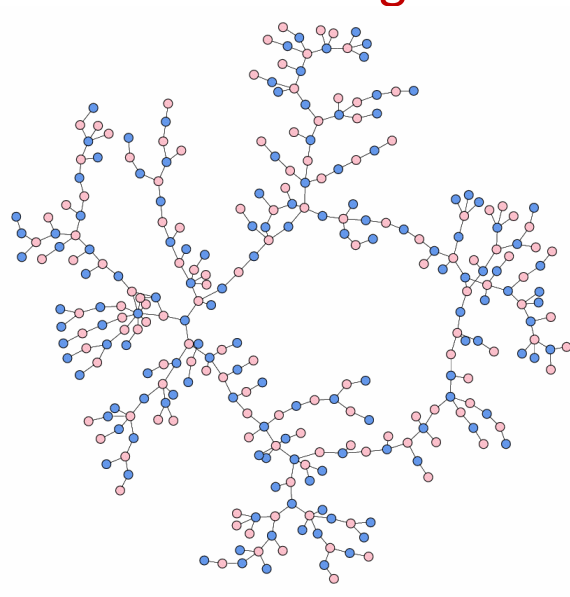


Y = marital status: never married?

Adult dataset, build 4 classifiers

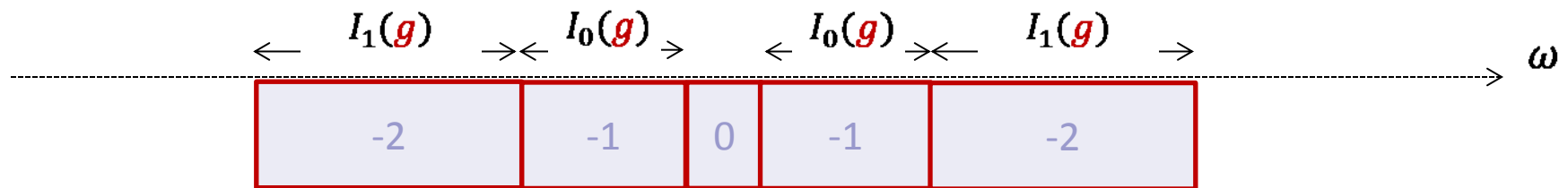
Example 2: Graph Data

- ◆ Releasing graph structured data remains a big challenge
 - Each individual (node) can have a big impact on graph structure
- ◆ **Current work** focuses on releasing graph statistics
 - Counts of small subgraphs like stars, triangles, cliques etc.
 - These counts are parameters for graph models
 - **Sensitivity of these counts is large**: one edge can change a lot



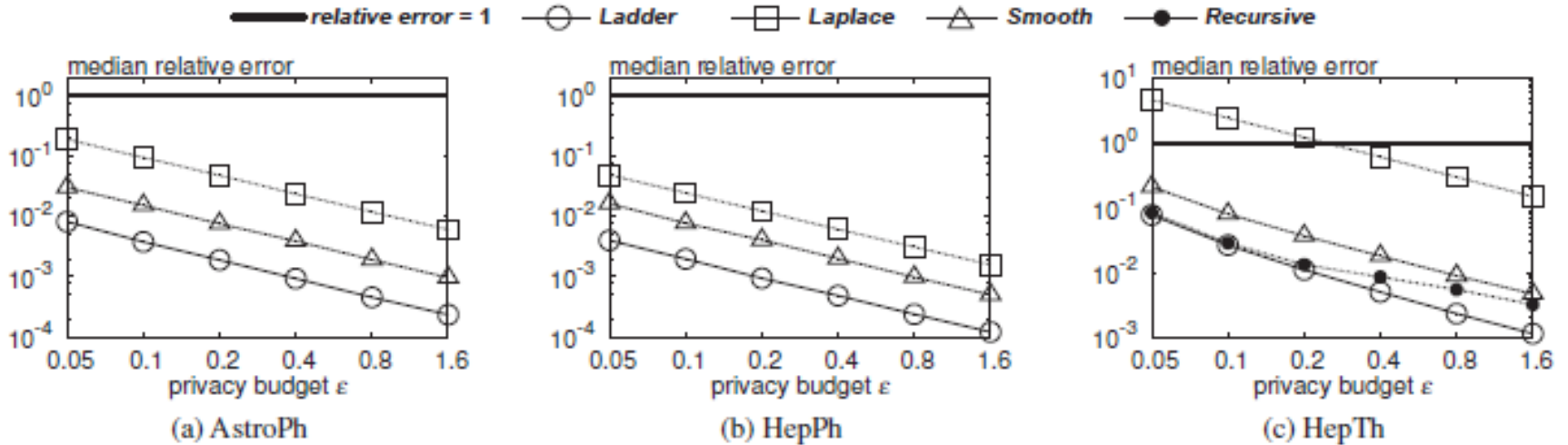
Staircase Mechanism [SIGMOD 15]

- ◆ **Our contribution:** dig deep into DP mechanisms for better results
 - Design a new “staircase mechanism” to release counts
 - Try to maximize likelihood of outputting correct answer
 - A carefully chosen function using ‘local sensitivity at distance d ’

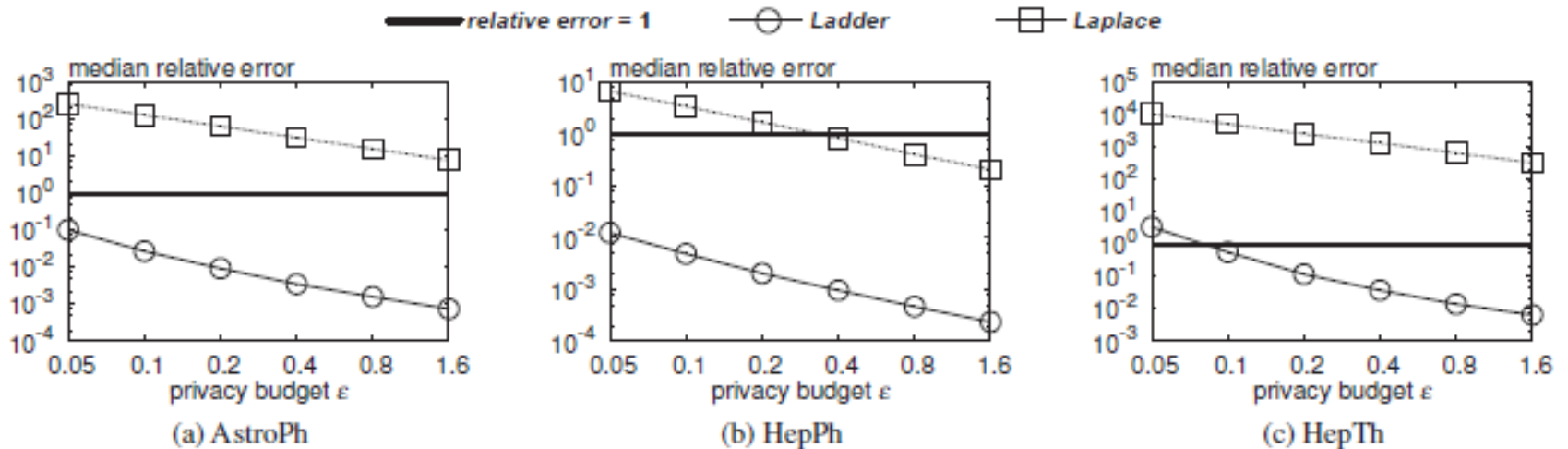


- ◆ Smaller relative error and faster results than prior work

Staircase experimental results

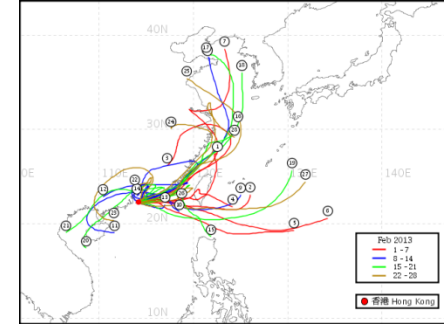


triangle counting



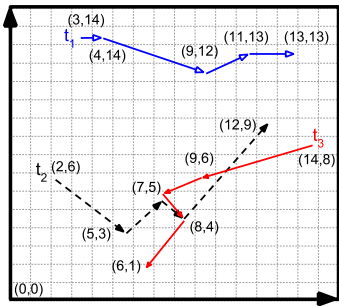
4-clique counting

Example 3: Trajectory Data

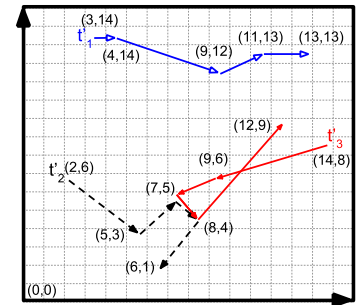


- ◆ More and more location and mobility data available
 - From GPS enabled devices, approximate location from wifi/phone
- ◆ Location and movements are **very sensitive!**
- ◆ Location and movements are **very identifying!**
 - Easy to identify ‘work’ and ‘home’ locations from traces
 - 4 random points identify 95% of individuals [Montjoye et al 2013]
- ◆ Aim for **Differentially Private Trajectories**
 - Find a model that works for trajectory data
 - Based on Markov models at multiple resolutions

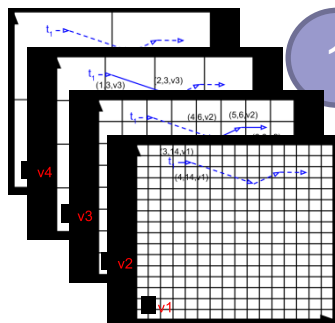
Original Trajectories



Synthetic Trajectories



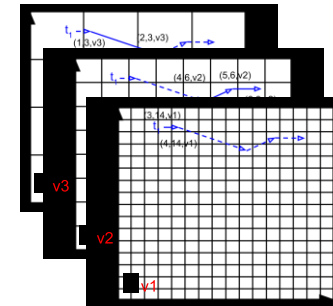
DPT System Overview



1 Hierarchical Reference System Mapping

6

Direction-weighted Sampling



2 Prefix Tree Construction



3 Model Selection

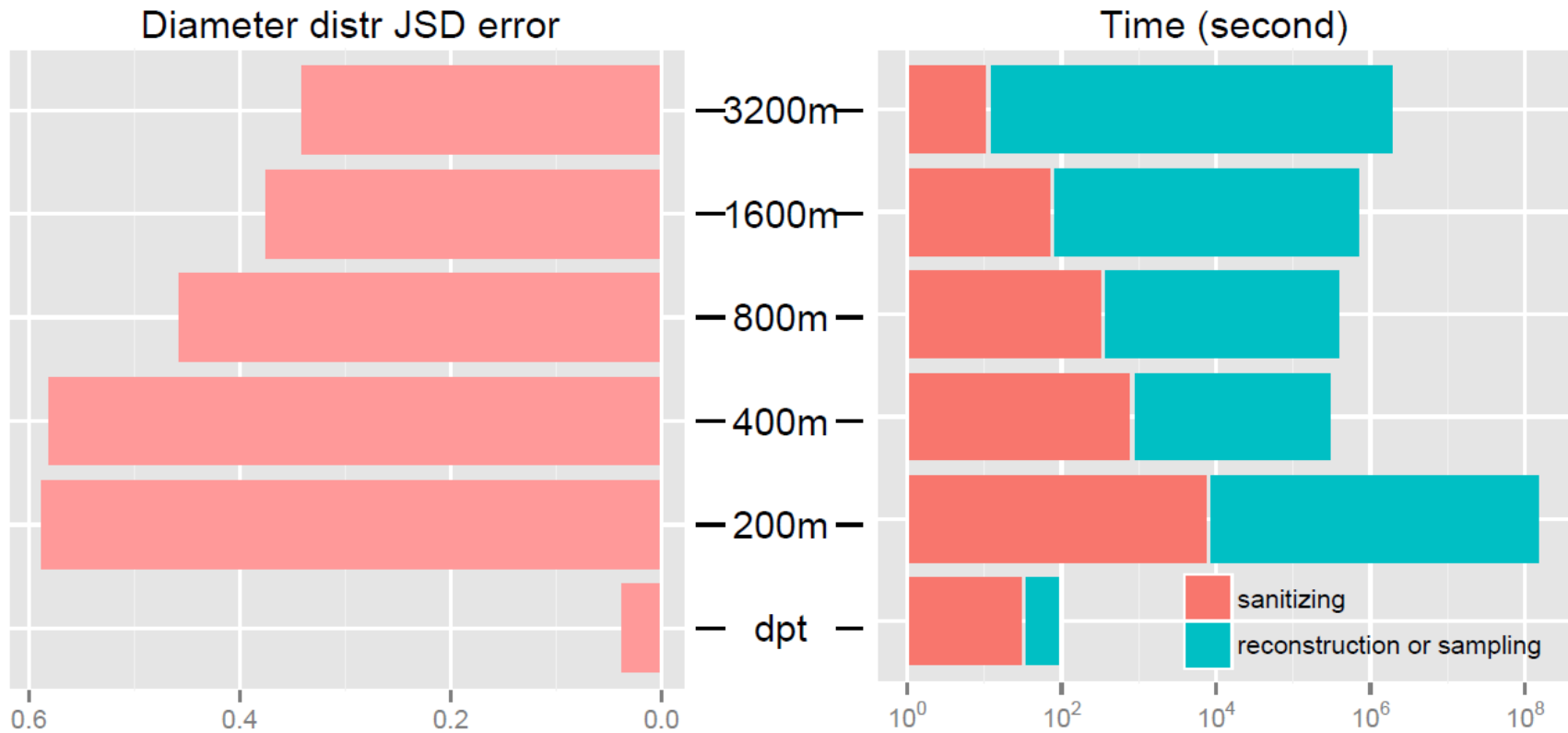


4 Noise Infusion



5 Adaptive Pruning

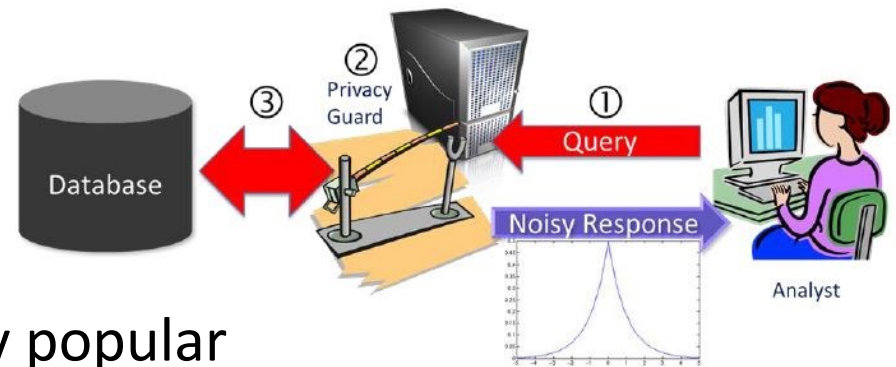
DPT results snapshot





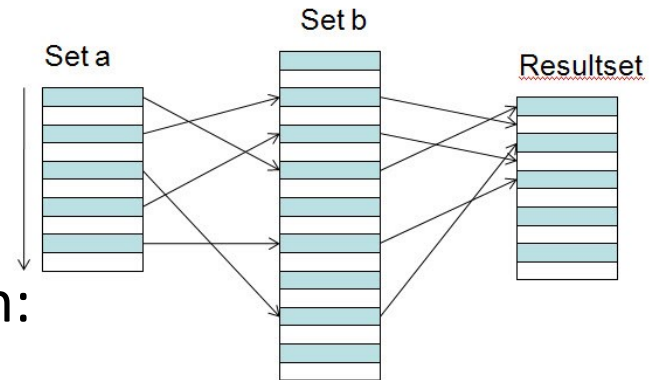
FUTURE: CHALLENGES AND PROBLEMS

DP Pros and Cons



- ◆ Differential privacy is currently popular
 - **Why?** Easy mechanisms and composition properties, deep theory
 - Proposed as an interactive mechanism, but easy to use for release
- ◆ Still some doubts and questions:
 - How to interpret ϵ ? How to set a value of ϵ ?
 - My answer: let $\epsilon \rightarrow \infty$ [let noise $\rightarrow 0$]
 - How robust is differential privacy in the wild?
 - It is possible to build an accurate classifier and make inferences
 - Sometimes the noise is just too high for utility: too much for some
- ◆ But alternate definitions have a high bar to entry...

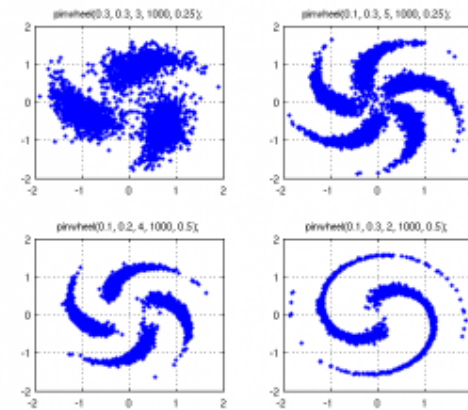
Challenge: private joins



- ◆ The following scenario occurs very often:
 - Companies **A** and **B** have data on people
 - They want to join their data on a unique identifier then remove it
 - They don't want the other to know their data
- ◆ Many possible approaches to solve in theory:
 - Some clever homomorphic cryptographic method
 - Introduce some trusted third party
 - Some careful use of hashing and salting
- ◆ **Current solution**: lots of wrangling between lawyers
 - **Open problem**: a practical solution to private joins?



Challenge: Better Synthetic Data



- ◆ Many past attempts to generate **synthetic data**
 - Avoids privacy concerns
 - But synthetic data based on a few parameters is unrealistic
- ◆ Aim for “**rich synthetic data**” instead
 - Make more use real data to instantiate models
- ◆ The gap between anonymized and synthetic data is eroding
 - Are these ultimately the same thing?
 - Varying in how much they depend on the original population
- ◆ Still need more work on effective synthetic data generation
 - Finding the right models, well-supported by the data

Challenge: Transition ideas to practice

- ◆ Many companies would like academics to work on their data



We have some great data for your team to look at!



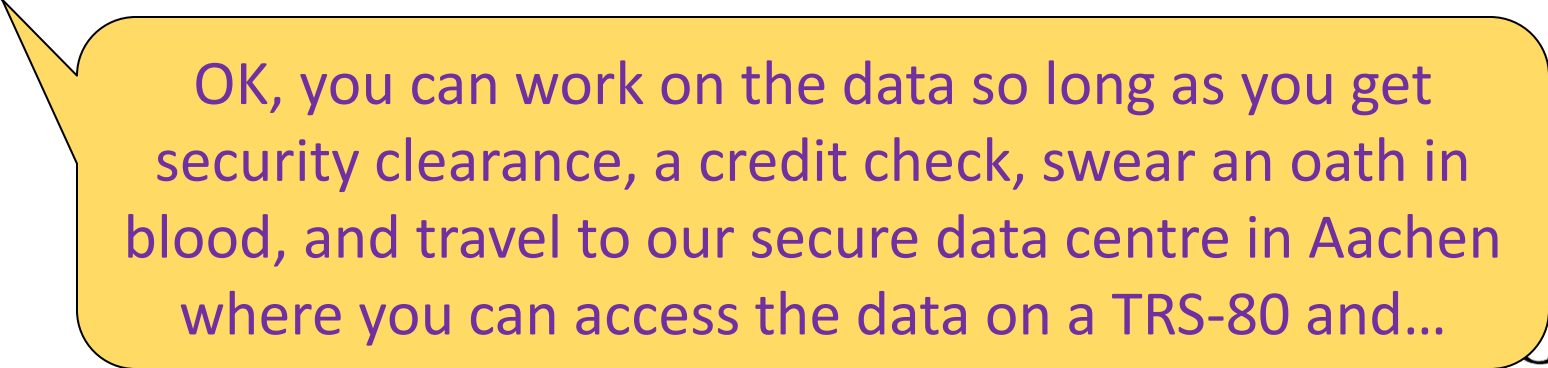
Thanks, but how are you going to deal with privacy issues?



It's fine, we can get you the data



... er, how's the release process going?

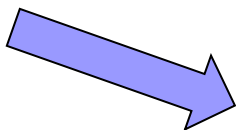
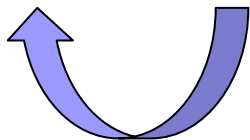
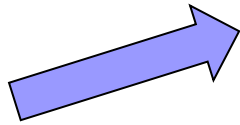


OK, you can work on the data so long as you get security clearance, a credit check, swear an oath in blood, and travel to our secure data centre in Aachen where you can access the data on a TRS-80 and...



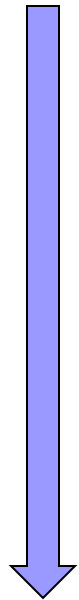
Dimensions to consider for data release

- ◆ How much **privacy** do we need?
- ◆ How much **utility** do we want from the anonymized data?
- ◆ How will data be accessed: as data feed, as data set, via API?

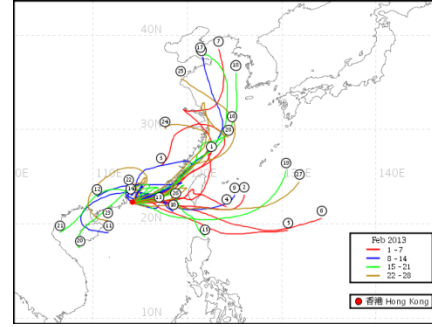


Who will use the data?

1. Trusted employees
Temporary employees
(students, contractors)
2. Outside contractors
Data purchasers
3. General Public



Summary



- ◆ Private data release is a **confounding problem!**
 - We haven't yet got it right consistently enough
 - The idea of “1 click privacy” is still a long way off
- ◆ Current privacy work gives some cause for **optimism**
 - Statistical privacy, safety in numbers, and robust models
- ◆ Lots of technical work left to do:
 - **Structured data**: graphs, movement patterns
 - **Unstructured data**: text, images, video?
 - Develop standards for (certain kinds of) data release



Joint work with Xi He, Divesh Srivastava, Magda Procopiuc,
Ashwin Machanavajhala, Xiaokui Xiao, Jun Zhang
Supported by Royal Society, European Commission

Database Research and Anonymization

- ◆ [SIGMOD] “Research papers will be judged ... through double-blind reviewing”
- ◆ [TODS] Authors need only apply 6 simple steps to blind their submission:
 1. Anonymize the title page
 2. Remove mention of funding sources and personal acknowledgments
 3. Anonymize references found in running prose that cite your papers
 4. Anonymize citations of submitted work in the bibliography
 5. Ambiguate statements on systems that identify an author
 6. Name your files with care, document properties are also anonymized
- ◆ How can this anonymization method be attacked?

