The confounding problem of private data release

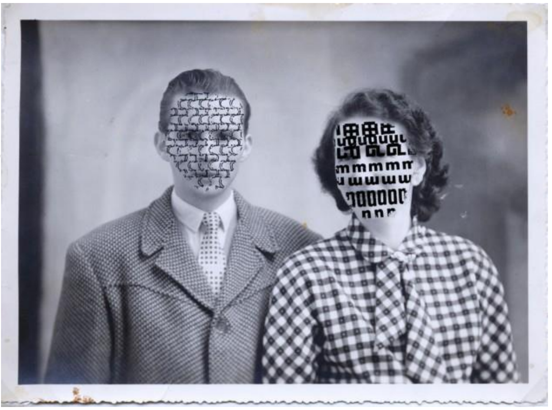
Graham Cormode

g.cormode@warwick.ac.uk

THE UNIVERSITY OF WARWICK

Big data, big problem?

- The big data meme has taken root
  - Organizations jumped on the bandwagon
  - Entered the public vocabulary
- But this data is mostly about individuals
  - Individuals want privacy for their data
  - How can researchers work on sensitive data?
- The easy answer: anonymize it and share
- The problem: we don’t know how to do this



Outline

- Why data anonymization is hard
- Differential privacy definition and examples
- Some snapshots of recent work
- A handful of new directions

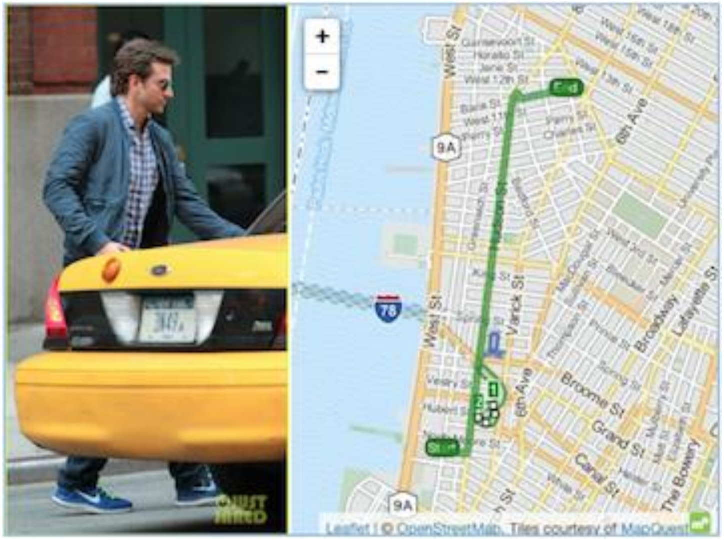
A moving example

NYC Taxi & Limousine Commission

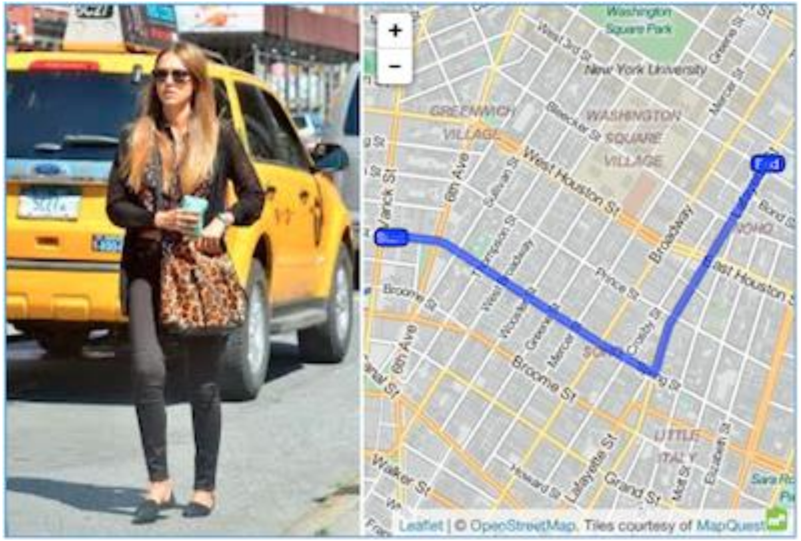
- NYC taxi and limousine commission released 2013 trip data
  - Contains start point, end point, timestamps, taxi id, fare, tip amount
  - 173 million trips “anonymized” to remove identifying information
- Problem: the anonymization was easily reversed
  - Anonymization was a simple hash of the identifiers
  - 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)

Sleuthing by Anthony Tockar while interning at Neustar

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...

Commonwealth of Massachusetts
Group Insurance Commission
*Your
Benefits
Connection*

NETFLIX

AOL

We need to solve this
data release problem...

Encryption 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 CEO and CTO off front page of newspapers
  - Simplify the process as much as possible: 1-click 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:

$$\Pr[\,K(D) \in S] \le e^{\varepsilon} \Pr[\,K(D') \in S]$$

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

Privacy with a coin toss

Perhaps the simplest possible DP algorithm

- Each user has a single private bit of information
  - Encoding e.g. political/sexual/religious preference, illness, etc.
- Toss a (biased) coin
  - With probability $p > ½$, report the true answer
  - With probability $1-p$, lie
- Collect the responses from a large number N of users
  - Can ‘unbias’ the estimate (if we know p) of the population fraction
  - The error in the estimate is proportional to $1/\sqrt{N}$
- Gives differential privacy with parameter $\ln (p/(1-p))$
  - Works well in theory, but would anyone ever use this?

Privacy in practice

- Differential privacy based on coin tossing is widely deployed
  - In Google Chrome browser, to collect browsing statistics
  - In Apple iOS and MacOS, to collect typing statistics
  - This yields deployments of over 100 million users
- The model where users apply differential privately and then aggregated is known as “Local Differential Privacy”
  - The alternative is to give data to a third party to aggregate
  - The coin tossing method is known as ‘randomized response’
- Local Differential privacy is state of the art in 2017: Randomized response invented in 1965: five decade lead time!

Going beyond 1 bit of data

- 1 bit can tell you a lot, but can we do more?
- Recent work: materializing marginal distributions
  - Each user has d bits of data (encoding sensitive data)
  - We are interested in the distribution of combinations of attributes

	Gender	Obese	High BP	Smoke	Disease
Alice	1	0	0	1	0
Bob	0	1	0	1	1
…					
Zayn	0	0	1	0	0

Gender/Obese	0	1
0	0.28	0.22
1	0.29	0.21

Disease/Smoke	0	1
0	0.55	0.15
1	0.10	0.20

Nail, meet hammer

- Could apply Randomized Reponse to each entry of each marginal
  - To give an overall guarantee of privacy, need to change p
  - The more bits released by a user, the closer p gets to ½ (noise)
- Need to design algorithms that minimize information per user
- First observation: a sampling trick
  - If we release n bits of information per user, the error is $n/\sqrt{N}$
  - If we sample 1 out of n bits, the error is $\sqrt{(n/N)}$
  - Quadratically better to sample than to share!

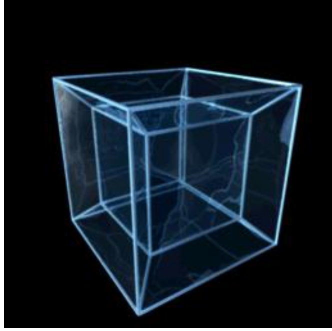
What to materialize?

Different approaches based on how information is revealed

1. We could reveal information about all marginals of size k
   - There are (d choose k) such marginals, of size 2^k each
2. Or we could reveal information about the full distribution
   - There are 2^d entries in the d-dimensional distribution
   - Then aggregate results here (obtaining additional error)

- Still using randomized response on each entry
  - Approach 1 (marginals): cost proportional to $2^{3k/2} d^{k/2}/\sqrt{N}$
  - Approach 2 (full): cost proportional to $2^{(d+k)/2}/\sqrt{N}$
- If k is small (say, 2), and d is large (say 10s), Approach 1 is better
  - But there's another approach to try…

Hadamard transform



Instead of materializing the data, we can transform it

- The Hadamard transform is the discrete Fourier transform for the binary hypercube
  - Very simple in practice

$$\begin{bmatrix} H^* & H^* \\ H^* & -H^* \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 \\ -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix} .$$

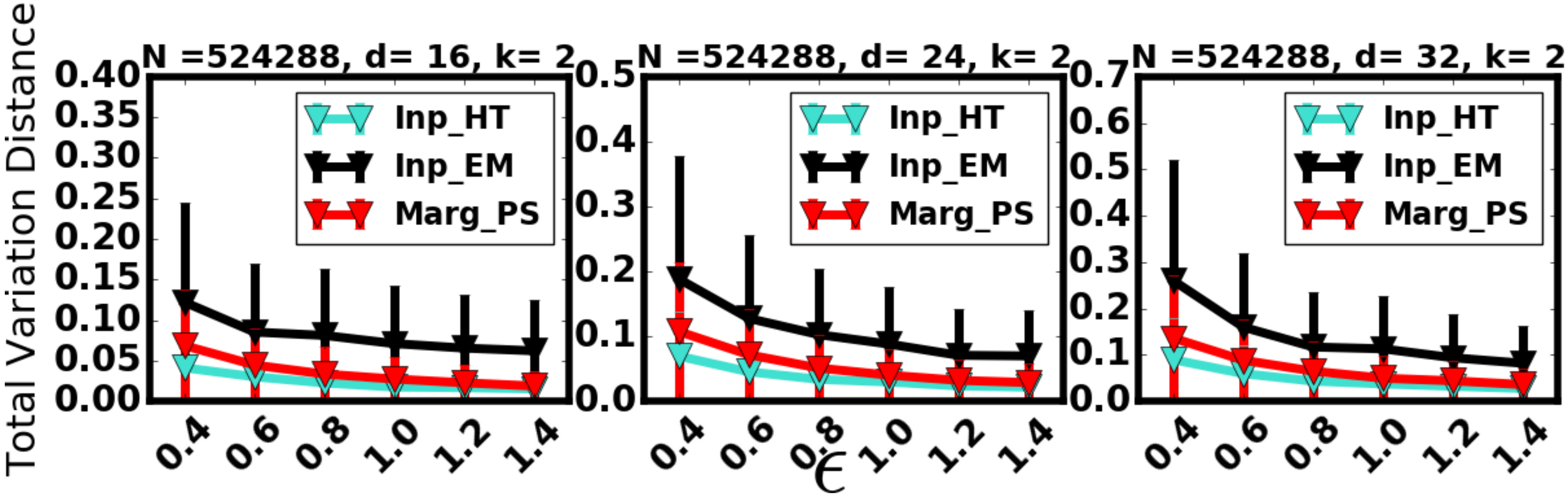
- Property 1: only (d choose k) coefficients are needed to build any k-way marginal
  - Reduces the amount of information to release
- Property 2: Hadamard transform is a linear transform
  - Can estimate global coefficients by sampling and averaging
- Yields error proportional to $2^{k/2}d^{k/2}/\sqrt{N}$
  - Better than both previous methods (in theory)

Outline of error bounds

How to prove these error bounds?

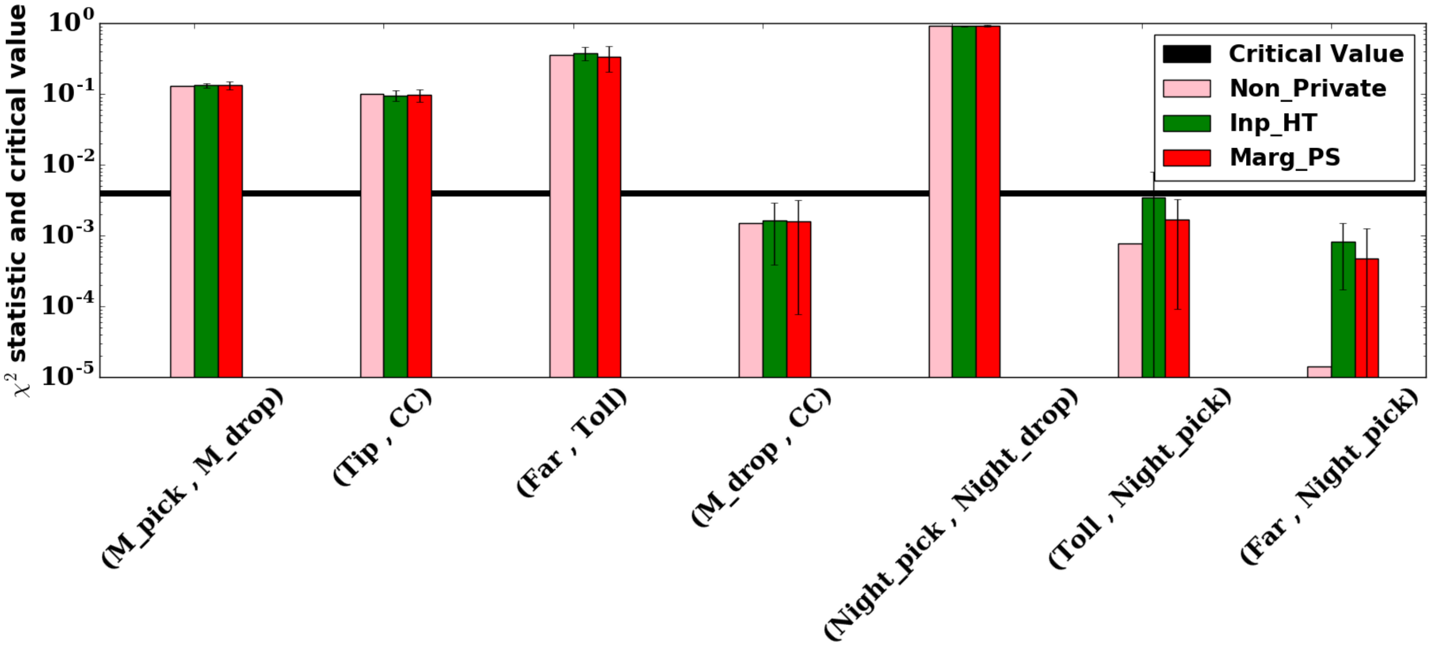
- Create a random variable X_i encoding the error from each user
  - Show that it is unbiased: $E[X_i]=0$, error is zero in expectation
- Compute a bound for its variance, $E[X_i^2]$ (including sampling)
- Use appropriate inequality to bound error of sum, $|\sum_{i=1}^{N} X_i|$
  - Bernstein or Hoeffding in equalities: error like $\sqrt{}(N/E[X_i^2])$
  - Typically, error in average of N goes as $1/\sqrt{}N$
- Possibly, second round of bounding error for further aggregation
  - E.g. first bound error to reconstruct full distribution, then error when aggregating to get a target marginal distribution

Empirical behaviour



- Compare three methods: Hadamard based (Inp_HT), marginal materialization (Marg_PS), Expectation maximization (Inp_EM)
- Measure sum of absolute error in materializing 2-way marginals
- N = 0.5M individuals, vary privacy parameter ε from 0.4 to 1.4

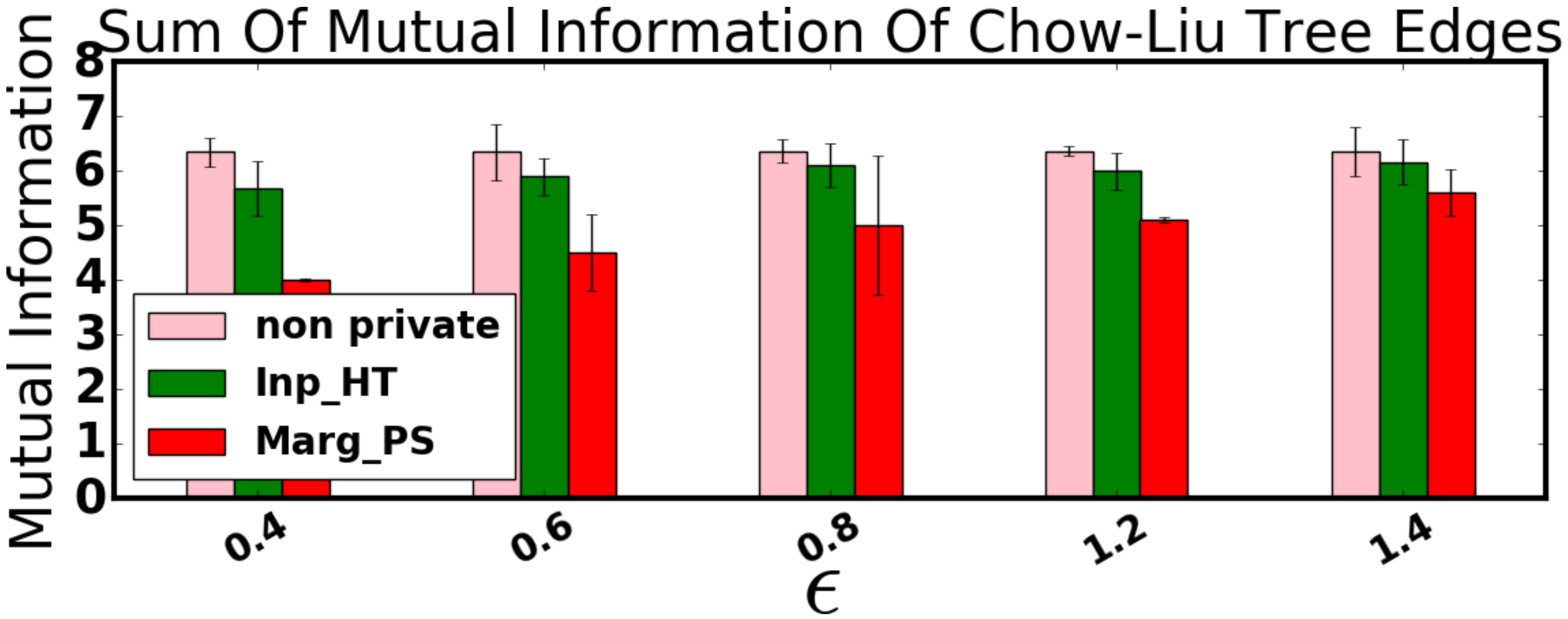
Applications – χ-squared test



- Anonymized, binarized NYC taxi data
- Compute χ-squared statistic to test correlation
- Want to be same side of the line as the non-private value!

THE UNIVERSITY OF WARWICK

Application – building a Bayesian model



- **Aim:** build the tree with highest mutual information (MI)
- Plot shows MI on the ground truth data for evaluation purposes

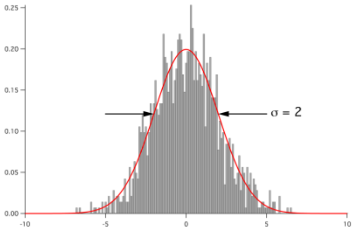
THE UNIVERSITY OF WARWICK

Centralized Differential Privacy

- There are a number of building blocks for centralized DP:
  - Geometric and Laplace mechanism for numeric functions
  - Exponential mechanism for sampling from arbitrary sets
    - Uses a user-supplied “quality function” for (input, output) pairs
- And “cement” to glue things together:
  - Parallel and sequential composition theorems
- With these blocks and cement, can build a lot
  - Many papers arrive from careful combination of these tools!

Differential privacy for data release

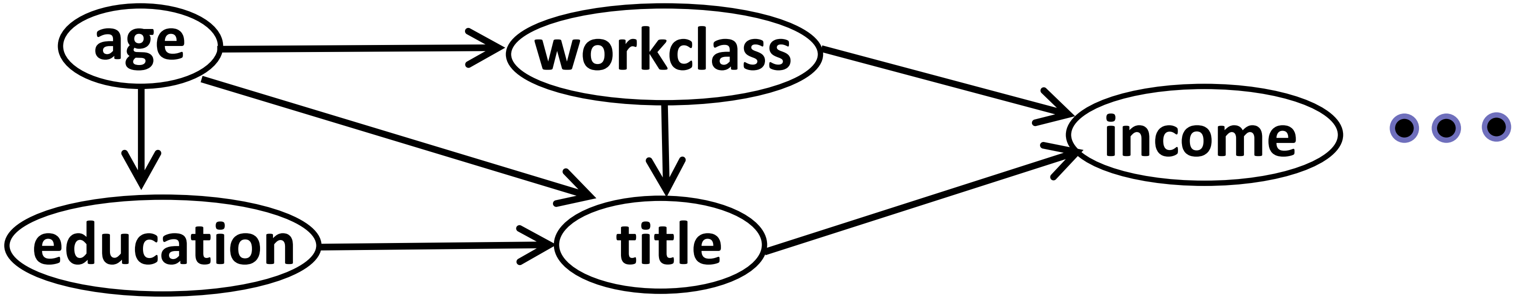
- Differential privacy is an attractive model for data release
  - Achieve a fairly robust statistical guarantee over outputs
- Problem: how to apply to data release where f(x) = x?
- General recipe: find a model for the data
  - Choose and release the model parameters under DP



- 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

Example: PrivBayes [TODS, 2017]

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



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

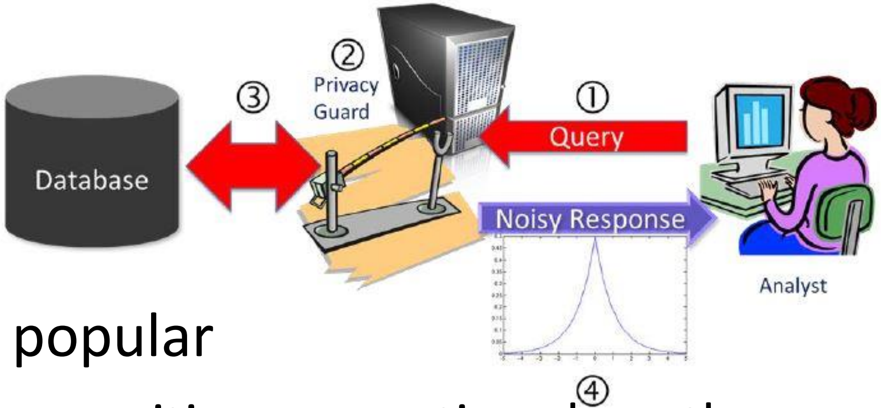
low-dimensional distributions: high signal-to-noise

THE UNIVERSITY OF WARWICK

PrivBayes

- **STEP 1:** Choose a suitable Bayesian Network BN
  - in a differentially private way
  - sample (via exponential mechanism) edges in the network
  - design surrogate quality function with low sensitivity
- **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)

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 ε → ∞ [let noise → 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 privacy definitions have a high bar to entry...

Challenge: Transition ideas to practice

- Many organizations 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…

Conclusions

- 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 opportunity for new work:
  - Designing optimal mechanisms for local differential privacy
  - Extend beyond simple counts and marginals
  - Structured data: graphs, movement patterns
  - Unstructured data: text, images, video?

Joint work with Divesh Srivastava (AT&T), Tejas Kulkarni (Warwick)

Supported by AT&T, Royal Society, European Commission