The confounding problem of private data release

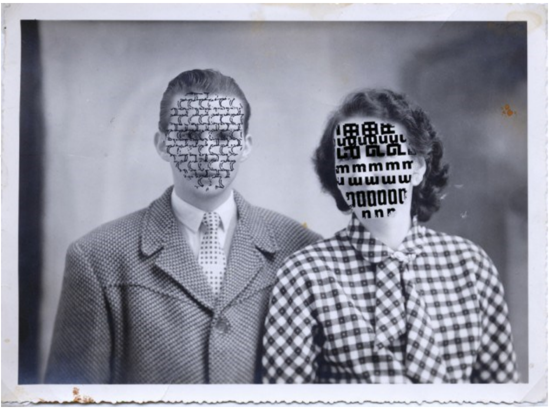
Graham Cormode

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
- Three snapshots of recent work
- A handful of new directions

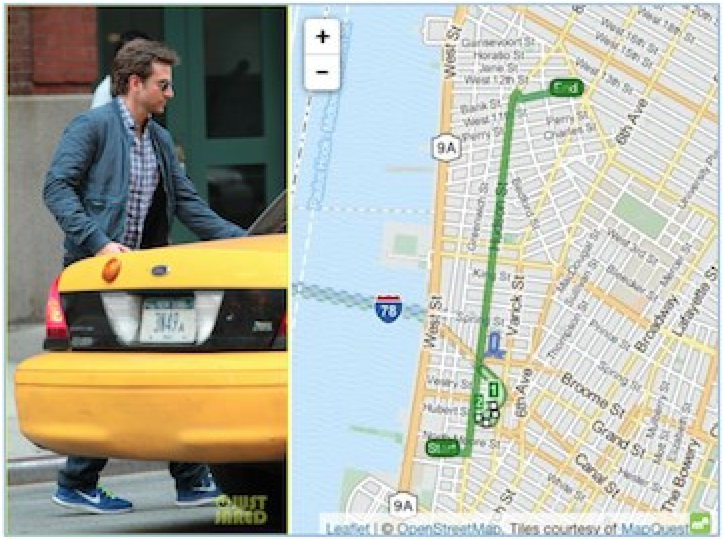
A recent data release example

NYC Taxi & Limousine Commission

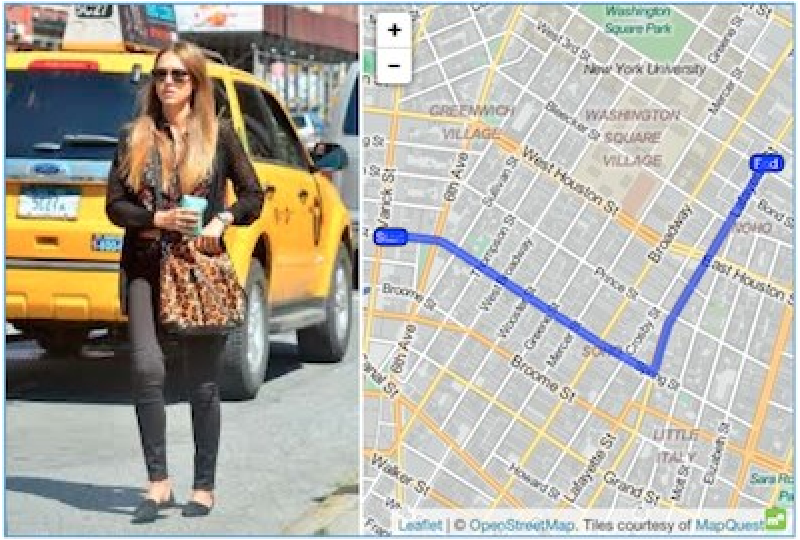
- NYC taxi and limousine commission released 2013 trip data
  - Contains start point, end point, timestamps, taxi id, fare, tip amount
  - 173 million trips "anonymized" to remove identifying information
- Problem: the anonymization was easily reversed
  - Anonymization was a simple hash of the identifiers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

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 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:

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

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

Achieving ε-Differential Privacy

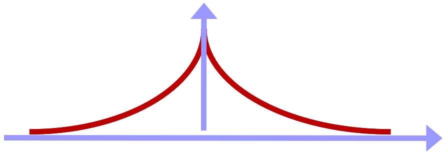
(Global) Sensitivity of publishing:

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

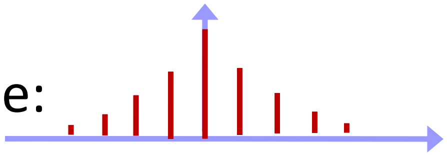
E.g., count individuals satisfying property P: one individual changing info affects answer by at most 1; hence $s = 1$

For every value that is output:

- Add Laplacian noise, $Lap(ε/s)$:



- 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

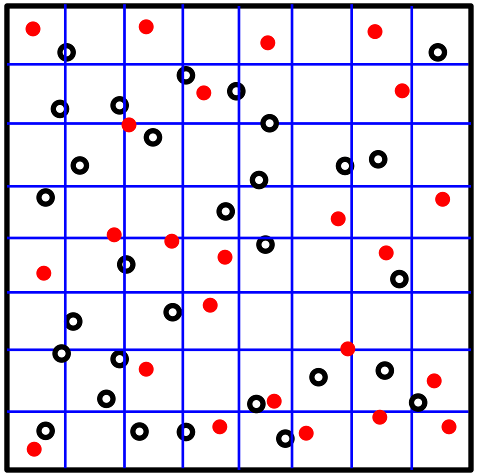
- (Sequential composition) If inputs overlap, result is $ε_1 + ε_2$ private
- (Parallel composition) If inputs disjoint, result is $\max(ε_1, ε_2)$ private

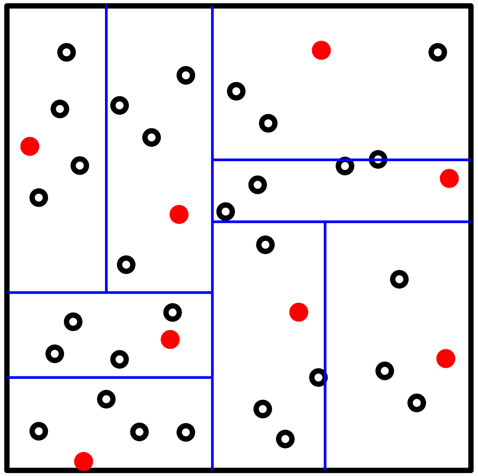
Technical Highlights

- There are a number of building blocks for 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!
- Useful fact: any post-processing of DP output remains DP
  - (so long as you don’t access the original data again)
  - Helps reason about privacy of data release processes

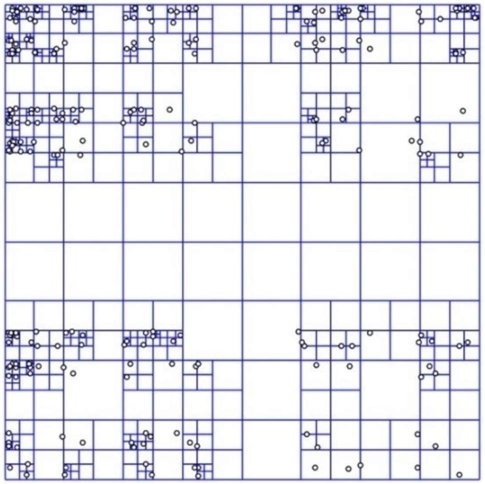
Case Study: Sparse Spatial Data

- Consider location data of many individuals
  - Some dense areas (towns and cities), some sparse (rural)
- Applying DP naively simply generates noise
  - lay down a fine grid, signal overwhelmed by noise
- Instead: compact regions with sufficient number of points

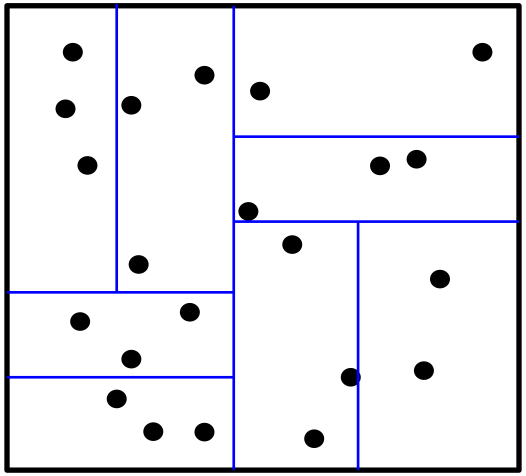




Private Spatial Decompositions (PSDs)



quadtree



kd-tree

- Build: adapt existing methods to have differential privacy
- Release: a private description of data distribution (in the form of bounding boxes and noisy counts)

Building a kd-tree

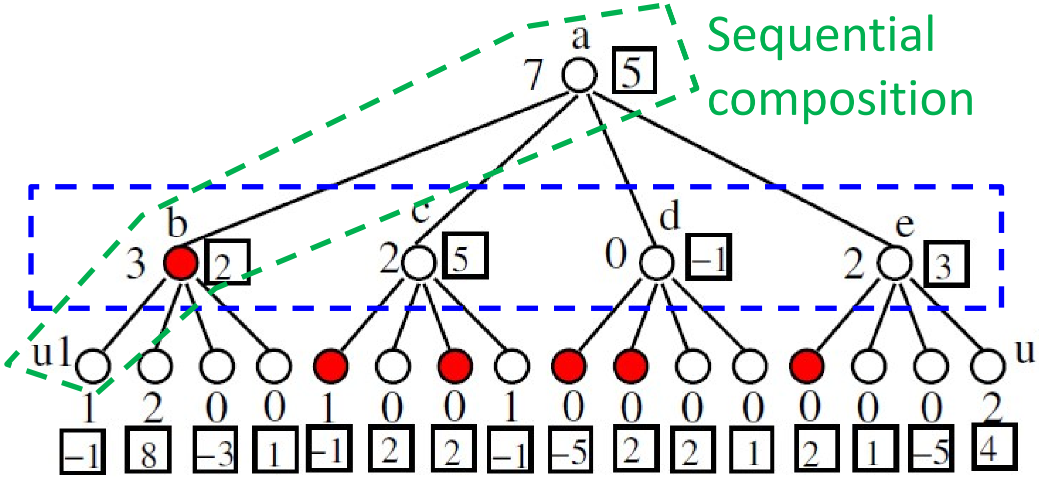
- Process to build a kd-tree
  - Input: data set
  - Choose dimension to split
  - Get median in this dimension
  - Create child nodes
  - Recurse until some stopping condition is met:
    - E.g. only 1 point remains in the current cell

Building a Private kd-tree

- Process to build a private kd-tree
  - Input: maximum height *h*, minimum leaf size *L*, data set
  - Choose dimension to split
  - Get (private) median in this dimension (exponential mechanism)
  - Create child nodes and add noise to the counts
  - Recurse until some stopping condition is met :
    - Max height is reached
    - Noisy count of this node less than *L*
    - Budget along the root-leaf path has used up
- The entire PSD satisfies DP by the composition property

Building PSDs – privacy budget allocation

- Data owner specifies a total budget ε reflecting the level of anonymization desired
- Budget is split between medians and counts
  - Tradeoff accuracy of division with accuracy of counts
- Budget is split across levels of the tree
  - Privacy budget used along any root-leaf path should total ε

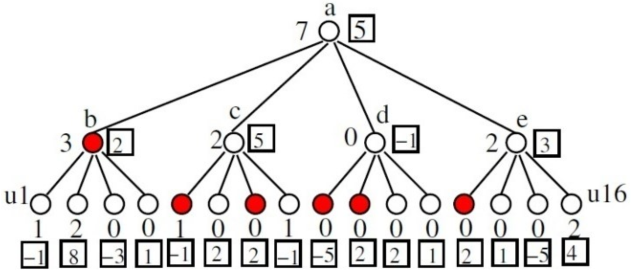


Sequential composition

Parallel composition

Privacy budget allocation

- How to set an ε_i for each level?
  - Compute the number of nodes touched by a ‘typical’ query
  - Minimize variance of such queries
  - Optimization: min $\sum_i 2^{h-i} / \varepsilon_i^2$ s.t. $\sum_i \varepsilon_i = \varepsilon$
  - Solved by $\varepsilon_i \propto (2^{(h-i)})^{1/3}\varepsilon$: more to leaves
  - Total error (variance) goes as $2^h/\varepsilon^2$



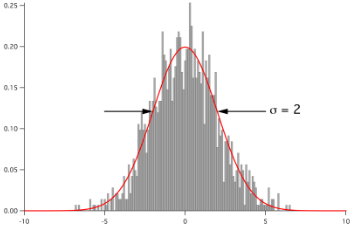
- Tradeoff between noise error and spatial uncertainty
  - Reducing h drops the noise error
  - But lower h increases the size of leaves, more uncertainty

Post-processing of noisy counts

- Can do additional post-processing of the noisy counts
  - To improve query accuracy and achieve consistency
- Intuition: we have count estimate for a node and for its children
  - Combine these independent estimates to get better accuracy
  - Make consistent with some true set of leaf counts
- Formulate as a linear system in n unknowns
  - Avoid explicitly solving the system
  - Expresses optimal estimate for node v in terms of estimates of ancestors and noisy counts in subtree of v
  - Use the tree-structure to solve in three passes over the tree
  - Linear time to find optimal, consistent estimates

Differential privacy for data release

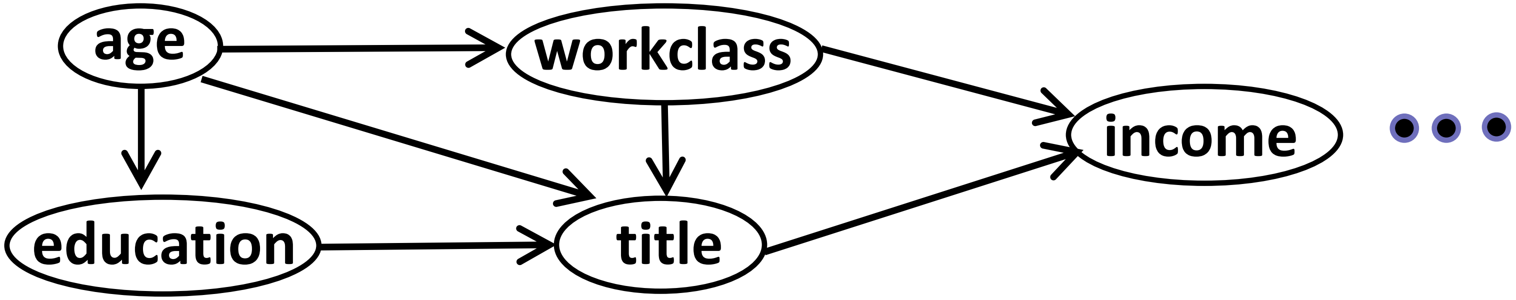
- Differential privacy is an attractive model for data release
  - Achieve a fairly robust statistical guarantee over outputs
- Problem: how to apply to data release where $f(x) = x$?
  - Trying to use global sensitivity does not work well
- General recipe: find a model for the data (e.g. PSDs)
  - Choose 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 [SIGMOD14]

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



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

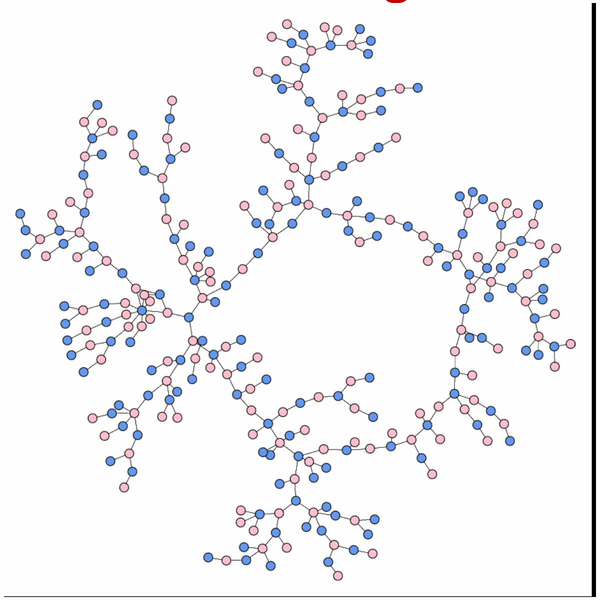
low-dimensional distributions: high signal-to-noise

PrivBayes (SIGMOD14)

- **STEP 1:** Choose a suitable Bayesian Network BN
  - in a differentially private way
  - sample (via exponential mechanism) edges in the network
  - 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)

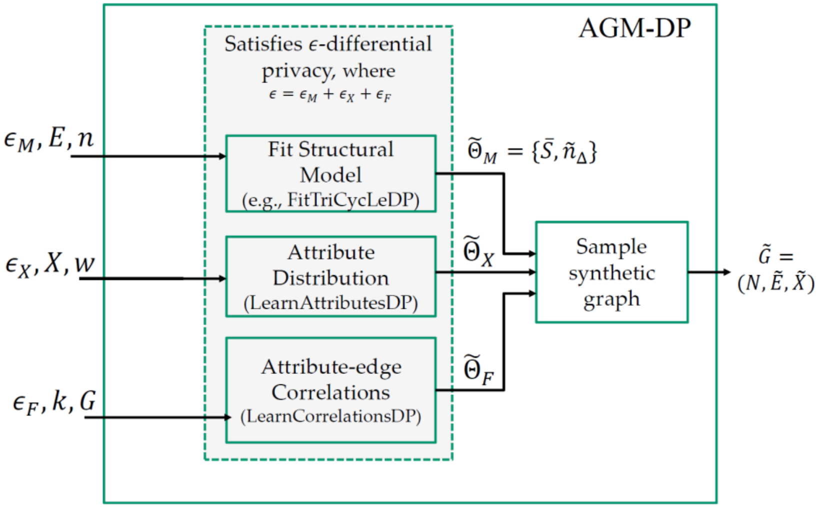
Example 2: Graph Data

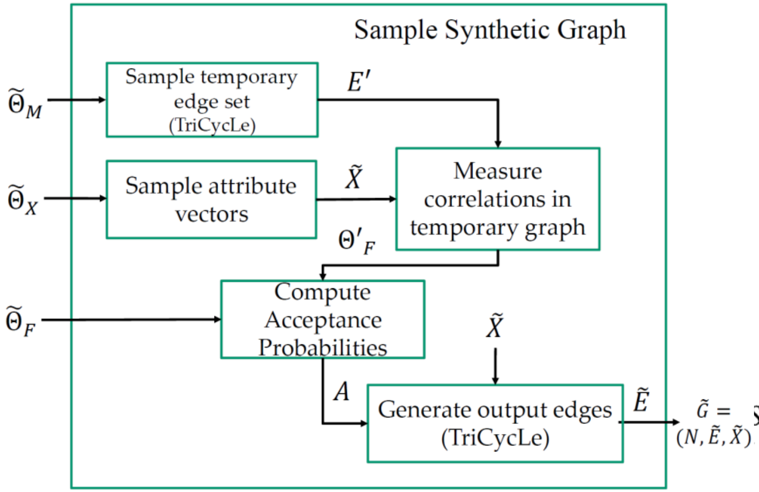
- Releasing graph structured data remains a big challenge
  - Each individual (node) can have a big impact on graph structure
- Current work focuses on releasing graph statistics
  - Counts of small subgraphs like stars, triangles, cliques etc.
  - These counts are parameters for graph models
  - Sensitivity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



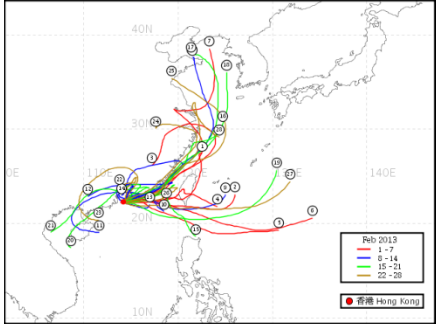
Attributed Graph Data [SIGMOD 16]

- Real graphs (e.g. social networks) have attributes
  - Different types of node, different types of edge
- Define graph models that have attribute distributions
  - Capture real graph structure e.g. number of triangles
- Learn parameters from input graphs (under differential privacy)
- Sample "realistic" graphs from the learned model


AGM-DP
Satisfies ϵ-differential privacy, where $\epsilon = \epsilon_M + \epsilon_X + \epsilon_F$
ϵ_M, E, n
ϵ_X, X, w
ϵ_F, k, G
Fit Structural Model (e.g., FitTriCycLeDP)
Attribute Distribution (LearnAttributesDP)
Attribute-edge Correlations (LearnCorrelationsDP)
$\widetilde{\Theta}_M = \{\bar{S}, \tilde{n}_\Delta\}$
$\widetilde{\Theta}_X$
$\widetilde{\Theta}_F$
Sample synthetic graph
$\tilde{G} = (N, \tilde{E}, \tilde{X})$

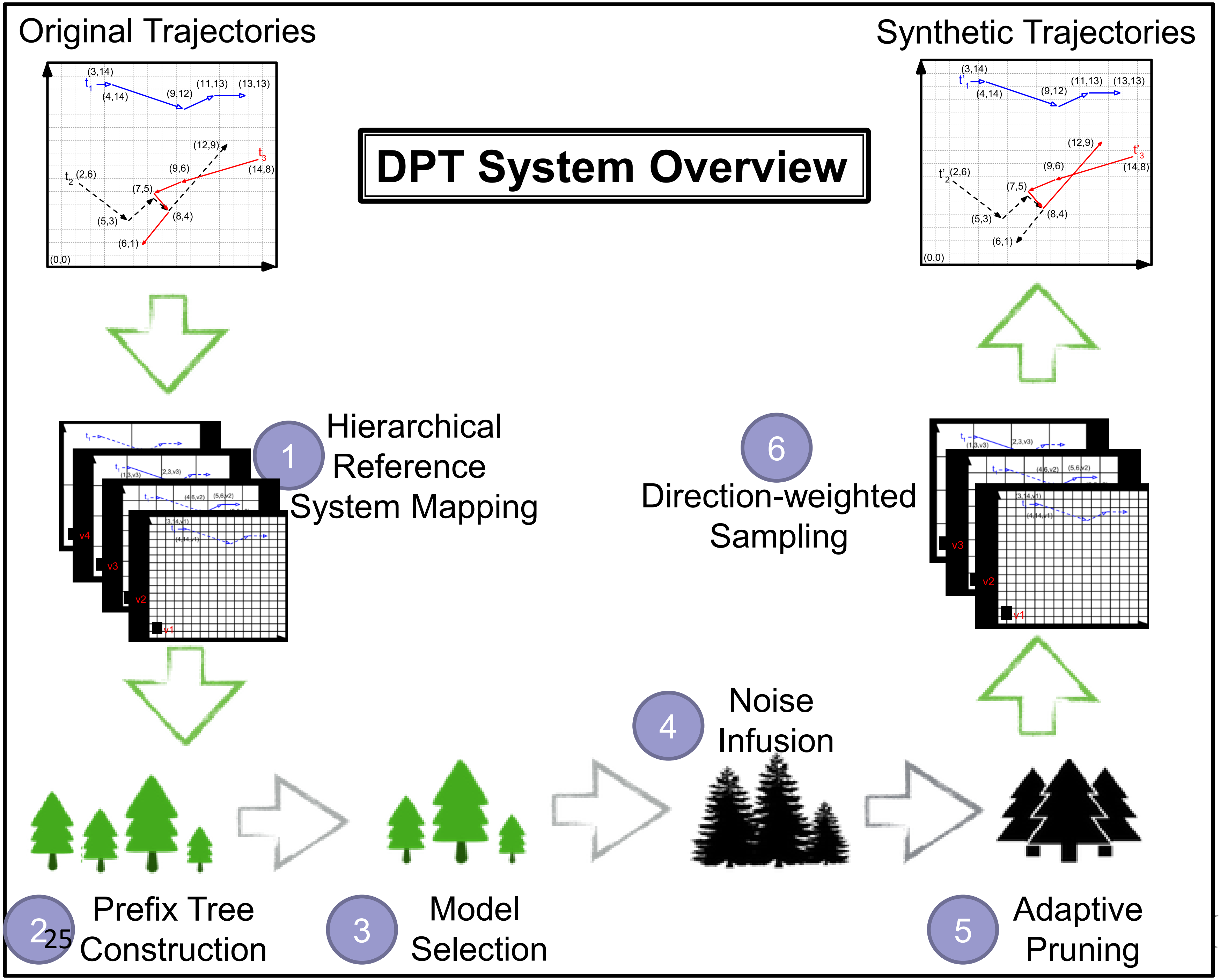

Sample Synthetic Graph
$\widetilde{\Theta}_M$
$\widetilde{\Theta}_X$
$\widetilde{\Theta}_F$
Sample temporary edge set (TriCycLe)
E'
Sample attribute vectors
$\tilde{X}$
Measure correlations in temporary graph
Θ'_F
Compute Acceptance Probabilities
$\tilde{X}$
A
Generate output edges (TriCycLe)
$\tilde{E}$
$\tilde{G} = (N, \tilde{E}, \tilde{X})$

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 [VLDB 15]
  - Find a model that works for trajectory data
  - Based on Markov models at multiple resolutions

DPT System Overview



Original Trajectories

Synthetic Trajectories

1 Hierarchical Reference System Mapping

2 Prefix Tree Construction

3 Model Selection

4 Noise Infusion

5 Adaptive Pruning

6 Direction-weighted Sampling

Local Differential Privacy

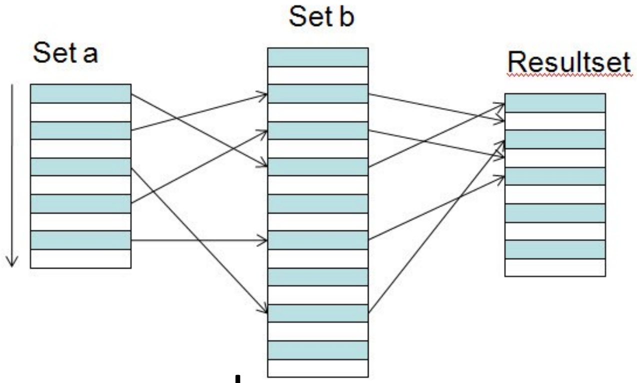
$$\begin{pmatrix} \mathrm{x} & x\alpha & x\alpha^2 & x\alpha^3 & \cdots & x\alpha^n \\ y\alpha & \mathrm{y} & y\alpha & y\alpha^2 & \cdots & y\alpha^{n-1} \\ y\alpha^2 & y\alpha & \mathrm{y} & y\alpha & \cdots & y\alpha^{n-2} \\ y\alpha^3 & y\alpha^2 & y\alpha & \mathrm{y} & \cdots & y\alpha^{n-3} \\ y\alpha^4 & y\alpha^3 & y\alpha^2 & y\alpha & \cdots & y\alpha^{n-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x\alpha^n & x\alpha^{n-1} & x\alpha^{n-2} & x\alpha^{n-3} & \cdots & \mathrm{x} \end{pmatrix}$$

- Data release assumes a trusted third party aggregator
  - What if I don’t want to trust a third party?
  - Back to crypto: fiddly secure multiparty computation protocols
- OR: run a DP algorithm with one participant for each user
  - Not as silly as it sounds: noise cancels over large groups
  - Implemented by Google and Apple (browsing/app statistics)
- Local Differential privacy state of the art in 2016: Randomized response (1965): five decade lead time!

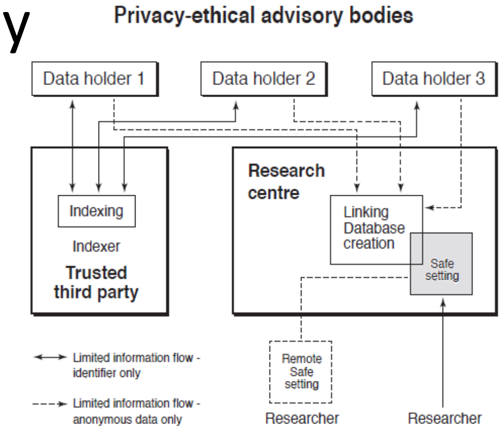
- Lots of opportunity for new work:
  - Designing optimal mechanisms for local differential privacy
  - Adapt to apply beyond simple counts

THE UNIVERSITY OF WARWICK

Trusted Third Parties



- The following scenario occurs very often:
  - Organizations A and B have collected 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
- Technical solutions may be possible, but complex
- Growing support for using a Trusted Third Party
  - Give data to TTP
  - They link the data sets, then remove ids
- ESRC’s Administrative Data Research Network:
  - Requests vetted for approval by experts
  - “There is a very small risk of ‘statistical disclosure’, when specific information from a de-identified data collection can be associated with a particular individual, household or business.”


Privacy-ethical advisory bodies
Data holder 1
Data holder 2
Data holder 3
Indexing
Indexer
Trusted third party
Research centre
Linking Database creation
Safe setting
Limited information flow - identifier only
Limited information flow - anonymous data only
Remote Safe setting
Researcher
Researcher

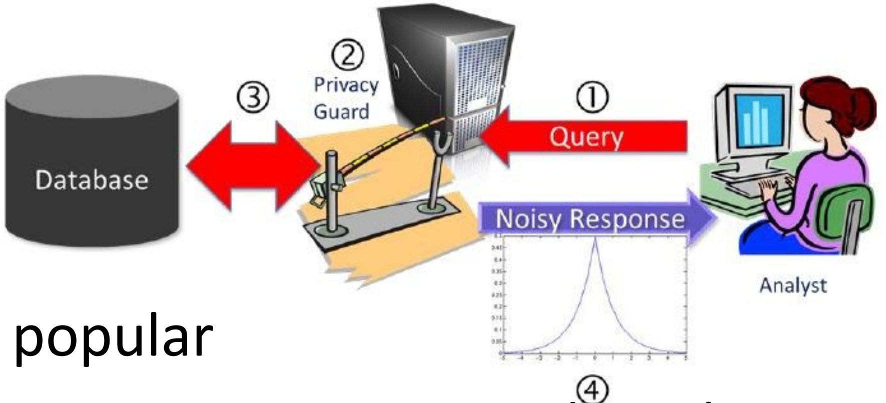
THE UNIVERSITY OF WARWICK

Care.data 2.0

- Care.data: 2014 effort to make all UK NHS data available to researchers (both academic and corporate)
  - National debate ensued, around poor communication of risks
  - Project delayed, seems to have ground to a halt
- 2016: DeepMind forms agreement with an NHS trust
  - 1.6M records shared for kidney disease study
  - Minor public comment
  - DeepMind promises to be very careful with the data
  - So that's OK then?

Google DeepMind

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!

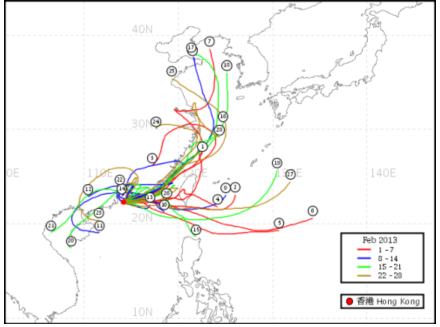
Thanks, but how are you going to deal with privacy issues?

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…

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, Tejas Kulkarni, Ashwin Machanavajjhala, Xiaokui Xiao, Jun Zhang, Zach Jorgensen

Supported by Royal Society, European Commission