

Getting Personal with Differential Privacy



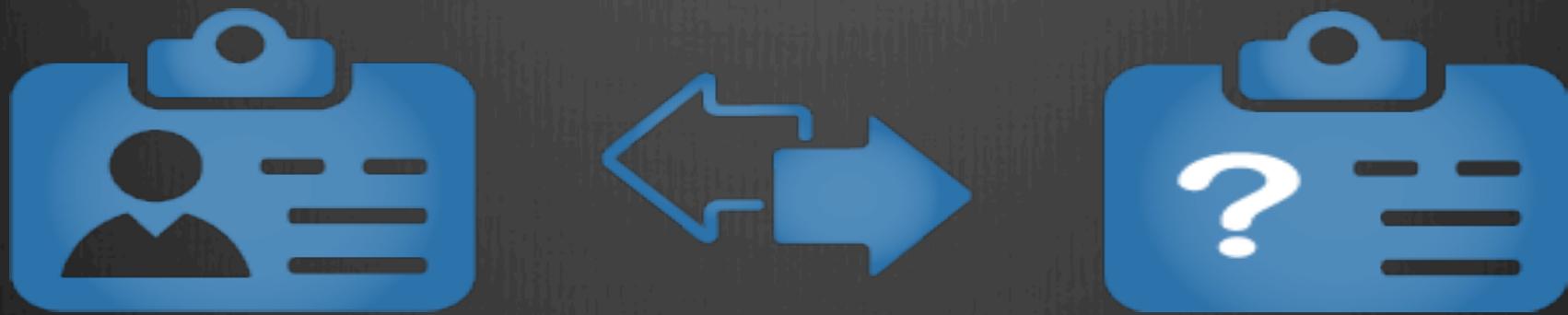
David Sands

Chalmers University of Technology
Sweden

Just how big are Dave's feet?



How Can We Ensure Privacy?



Anonymisation

Massachusetts
Group Insurance
Commission
released
"anonymized"
health records on
state employees



Photo:media.masslive.com

Anonymisation Fail

Netflix released viewer data for half a million subscribers a \$1M competition to build the best movie-recommender system

The screenshot shows the Netflix Prize Leaderboard page. At the top, the Netflix logo is visible above a yellow banner with the text "Netflix Prize". Below the banner, there is a navigation bar with links for "Home", "Rules", "Leaderboard", and "Update". The main title "Leaderboard" is displayed in large blue text. A sub-instruction "Showing Test Score. [Click here to show quiz score](#)" is present. A dropdown menu allows users to "Display top 20 leaders". The table below lists the top three teams:

Rank	Team Name	Best Test Score	% Improve
1	BellKor's Pragmatic Chaos	0.8567	10.06
2	The Ensemble	0.8567	10.06
3	Grand Prize Team	0.8582	9.90

Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos

Candidate definition of Absolute Privacy [Dalenius '77]:

Whatever you learn about an individual from a database could already have been learned without access to the database





The 1991 Romanian Mititei Survey



61% Pork
12% Beef

8.94 cm

3.6 cm

Does the 1991 “Mici” Survey Protect Dave’s Privacy?

“Mici” is a survey of Italian households conducted by the National Institute of Statistics (Istat) every five years since 1991. It is a large-scale survey that provides information on various socio-economic and demographic variables. One of the key features of the survey is its strict adherence to privacy principles, particularly regarding the protection of individual respondents’ data.

The survey uses a complex sampling design to ensure that the sample is representative of the Italian population. It employs a two-stage sampling process, where first, regions are selected, and then households are chosen within each region. This ensures that the survey covers all parts of the country and includes a diverse range of households.

One of the most important aspects of the survey is its commitment to data protection. All data collected is anonymized, meaning that individual respondents cannot be identified from the survey results. This is achieved through the use of unique household identifiers and the application of strict data handling procedures.

The survey also includes a variety of questions on topics such as household composition, income, education, and health. These questions are designed to provide a comprehensive picture of the living conditions and social circumstances of Italian households.

In conclusion, the 1991 “Mici” Survey is a well-designed and privacy-conscious survey that provides valuable information on Italian households. Its strict adherence to data protection principles ensures that individual respondents’ privacy is protected, making it a valuable tool for understanding the social and economic conditions of the Italian population.

Dave's feet are 3x longer than
the average 1991 mici



Anonymisation Cannot Guarantee Privacy

On the Difficulties of Disclosure Prevention
in Statistical Databases or The Case for
Differential Privacy, [Dwork & Naor 2010]

Differential Privacy

[Dwork & McSherry '06]

A **quantified** definition of privacy for a **noisy** statistical query:

quantify the *difference* in what might be learned about any individual from a database with or without said individual

Privacy Preserving Database Queries

- What is privacy for database queries?
 1. Introducing Differential Privacy
- How to build tools which make it easy to program data analyses while respecting privacy
 2. Building-blocks for DP mechanisms
- Outline of our approach:
 3. Personalised Differential Privacy

Differential Privacy

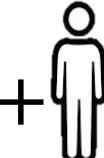
- A measure of the extent to which anyone can blend into the crowd
- A measure of the plausible deniability of the claim: “I’m not even in that database”

ϵ -Differentially Private query:

For any dataset  and any individual ,

the chance of getting answer A on 

vs

the chance of getting answer A on  + 

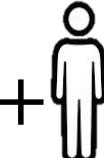
differ by at most a factor of $\exp(\epsilon)$

ϵ -Differentially Private query:

For any dataset  and any individual ,

the chance of getting answer A on 

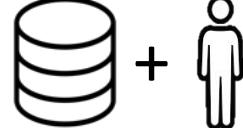
vs

the chance of getting answer A on  + 

differ by at most a factor of $1 \pm \epsilon$

Differential Privacy

$D =$ 

$D' =$ 

$\Pr[Q(D) = A]$

$\Pr[Q(D') = A]$

if you join the database, no outcome will change its probability by much

A

ratio bounded

Designing Differentially Private Mechanisms

This talk:

Intro to DP

+

dynamic enforcement method for
DP by information flow tracking

No statistical knowledge required!

Building DP Mechanisms

FOR

DUMMIES®

Just like using
LEGO (TM)!

*A Reference
for the
Rest of Us!*



Building Blocks for Differential Privacy

Compositionality principles make it easier to build differentially private mechanisms from components



Sequential Composition

An ϵ_1 -DP query, followed by

an ϵ_2 -DP is $(\epsilon_1 + \epsilon_2)$ -DP

[McSherry]

Holds even if Q_2 is chosen using the result of Q_1

Sensitivity (stability)

$$D =$$



$$D' =$$



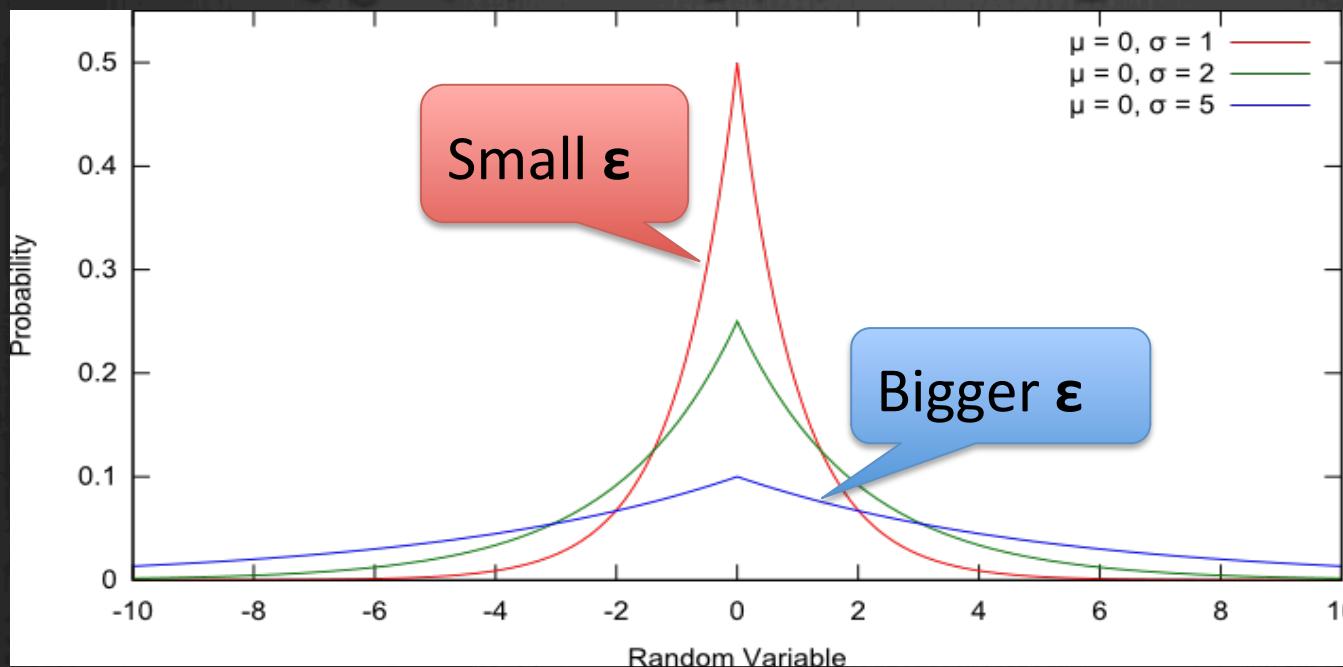
A function F has sensitivity S if $F(D)$ and $F(D')$ are different by at most (size) S

- count
- select males
- sum

Private Query = Query + Noise

If Q has sensitivity s then we can compute an ϵ -differentially private version of Q :

$$Q_\epsilon(x) = Q(x) + \text{Laplace}(s/\epsilon)$$

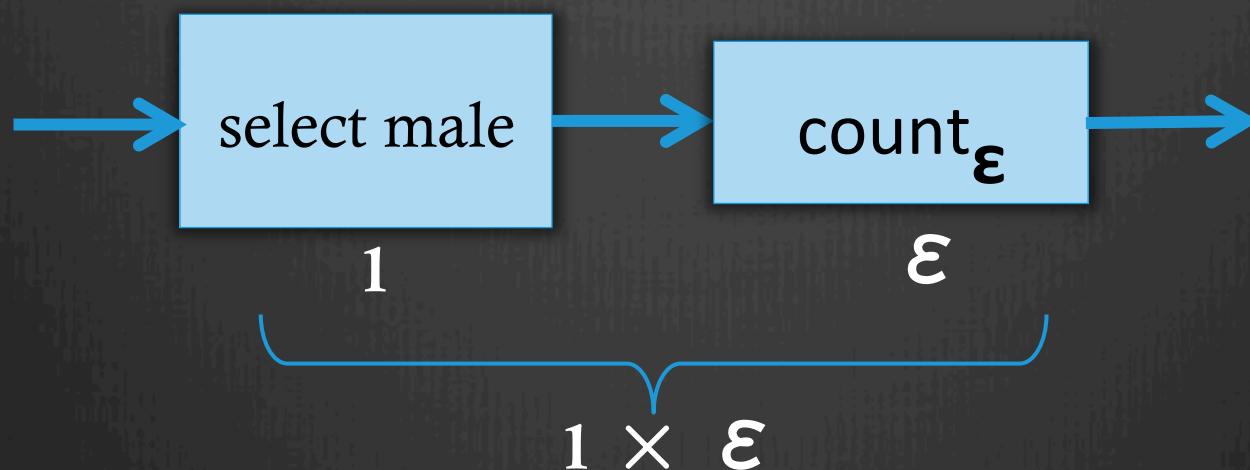


Laplace
distribution

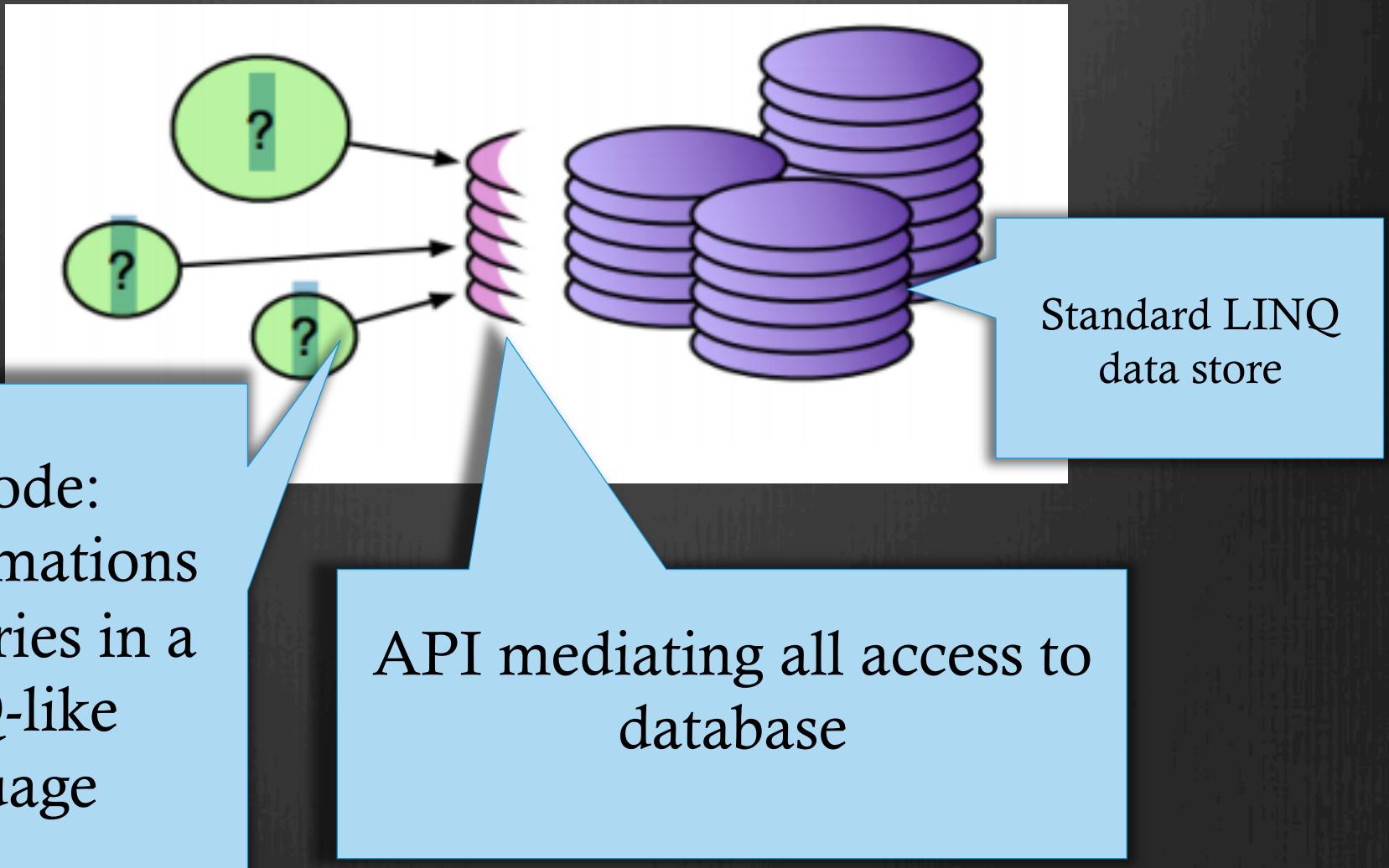
Sensitivity Composition

T has sensitivity s and Q is ϵ -DP then

$Q \circ T$ is $(s \times \epsilon)$ -DP

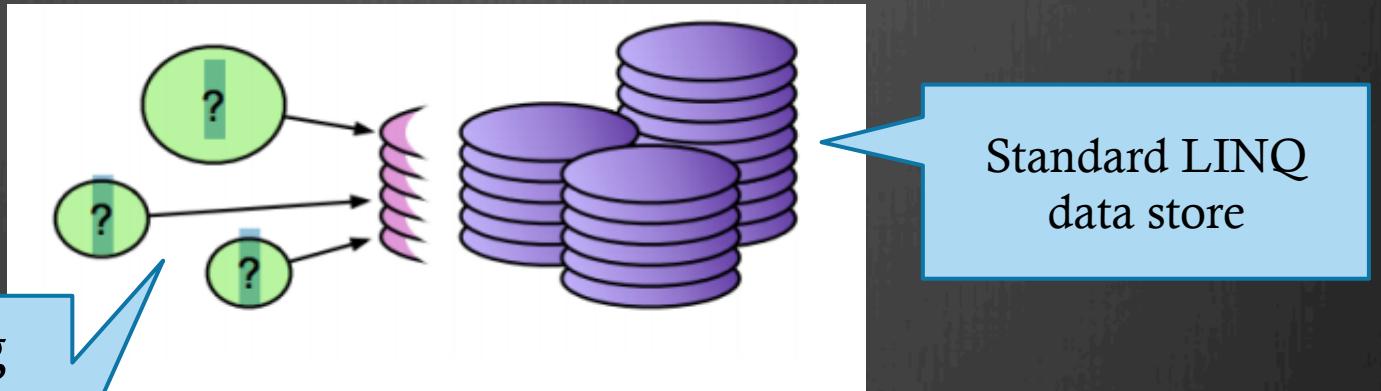


PINQ [McSherry]



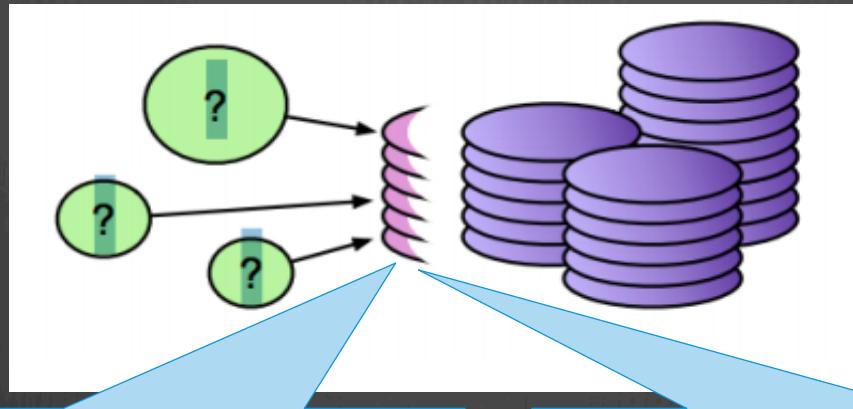
PINQ [McSherry]

C# code using
Transformations
and Queries in a
LINQ-like
language



```
var data = new PINQueryable<SearchRecord>(... ...);  
  
var users = from record in data  
            where record.Query == argv[0]  
            groupby record.IPAddress;  
  
Console.WriteLine(argv[0] + ":" + users.Count(0.1));
```

PINQ [McSherry]



Data

- A Global Privacy Budget
- The sensitivity of each intermediate database



Bookkeeping

- Deduct $\epsilon \times s$ from budget if ϵ query is applied to a table with sensitivity s
- Deny query whenever the budget is insufficient

Problem 1: Wasteful Global Budget



Detailed, multi-dimensional survey
of people with blood type *AB-*
negative

Budget Exhausted

Marketing study of all adults

Problem 2: Continuous Data



Detailed, multi-dimensional survey
of people with blood type *AB-*
negative

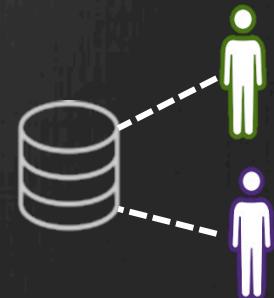
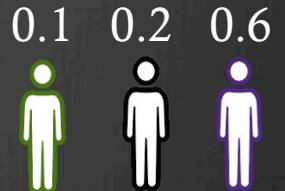


New data input to the database

Personalised Differential Privacy (PDP)

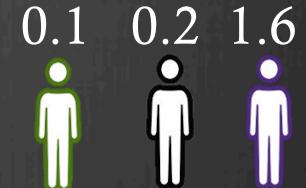
1. Generalise DP: each individual has their own personal “epsilon”
2. Show that PDP has its own composition principles
3. Implement PDP by tracking exact *provenance* of every record

[Ebadi, DS, Schneider, POPL 2015]



Personal (Big-Epsilon) Differential Privacy

Let E be a function from individual records to $\mathbb{R}^{\geq 0}$

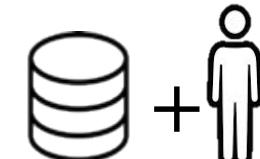


Query Q provides E -Differential Privacy

if for all  and all 

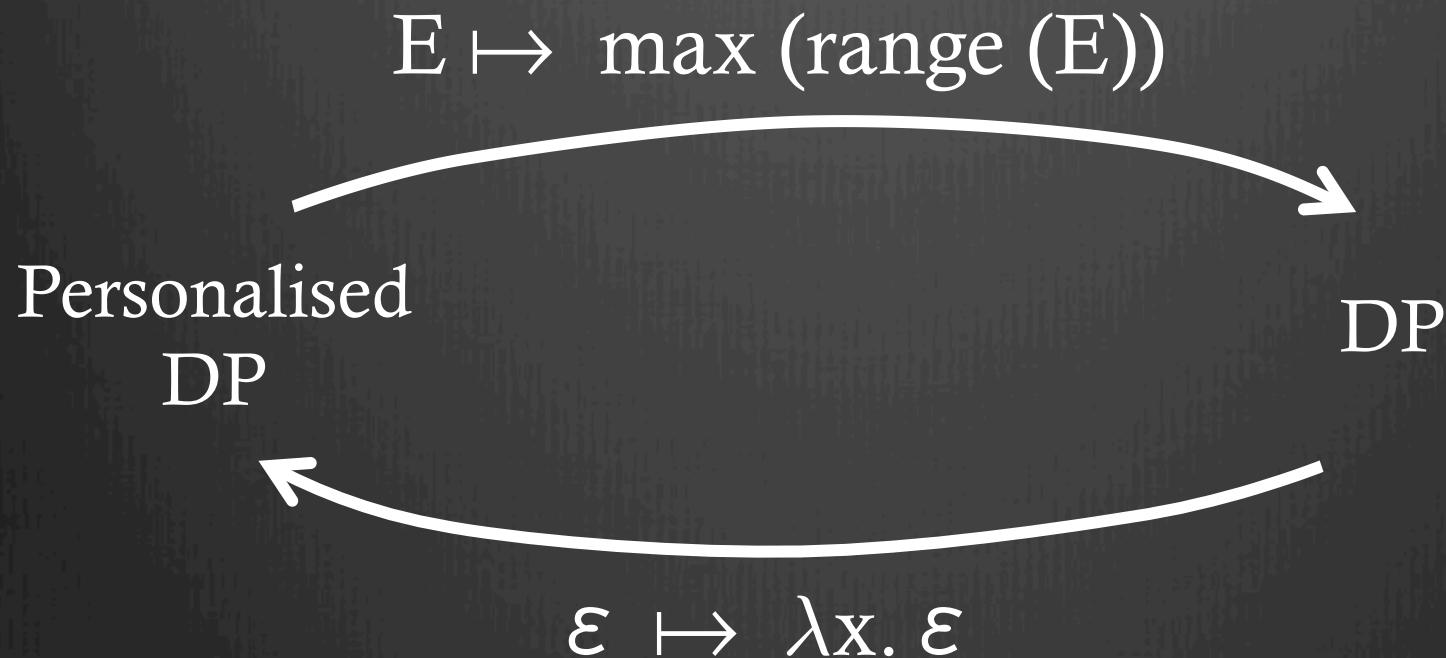
the chance of getting answer A on 

vs

the chance of getting answer A on 

differ by at most a factor of $1 \pm E(\text{person})$

PDP generalises DP



If Q is E -DP then Q is ε -DP for $\varepsilon = \max(\text{range}(E))$

PDP composition principles

Sequential composition

An E_1 -DP query, followed by an E_2 -DP query
is E -DP

$$\text{where } E(y) = E_1(y) + E_2(y)$$

PDP composition principles

Sensitivity composition

If Q is E -DP then $Q \circ F$ is E' -DP

where $E'(z) = \text{sensitivity}(F) \times E(z)$

PDP composition principles

“Computing the (noisy) average income of adult smokers is 0-differentially private for Jimmy, aged 10.”

Selection: select_P removes elements not in P

Selection composition principle:

Q is E -DP then $Q \circ \text{select}_P$ is E' -DP

where $E'(x) = \begin{cases} E(x) & \text{if } x \in P \\ 0 & \text{else} \end{cases}$

Union-preserving functions

$$F(A \cup B) = F(A) \cup F(B)$$

E.g. select, project, rename, map...

Union-preserving functions

If Q is ε -DP then $Q \circ F$ is E -DP,

where $E(x) = \varepsilon \times \text{size}(F\{x\})$

F magnifies the privacy cost of Q for Bob by $|F\{\text{Bob}\}|$

Provenance for Personalised Differential Privacy



m-w.com

A screenshot of the Merriam-Webster dictionary website. At the top, there are four tabs: 'Dictionary' (highlighted in blue), 'Thesaurus', 'Medical', and 'Encyclo.'. Below the tabs, the word 'provenance' is typed into a search bar. To the right of the search bar is a red square icon containing a white magnifying glass. The background of the search interface is blue.

provenance

Save
★

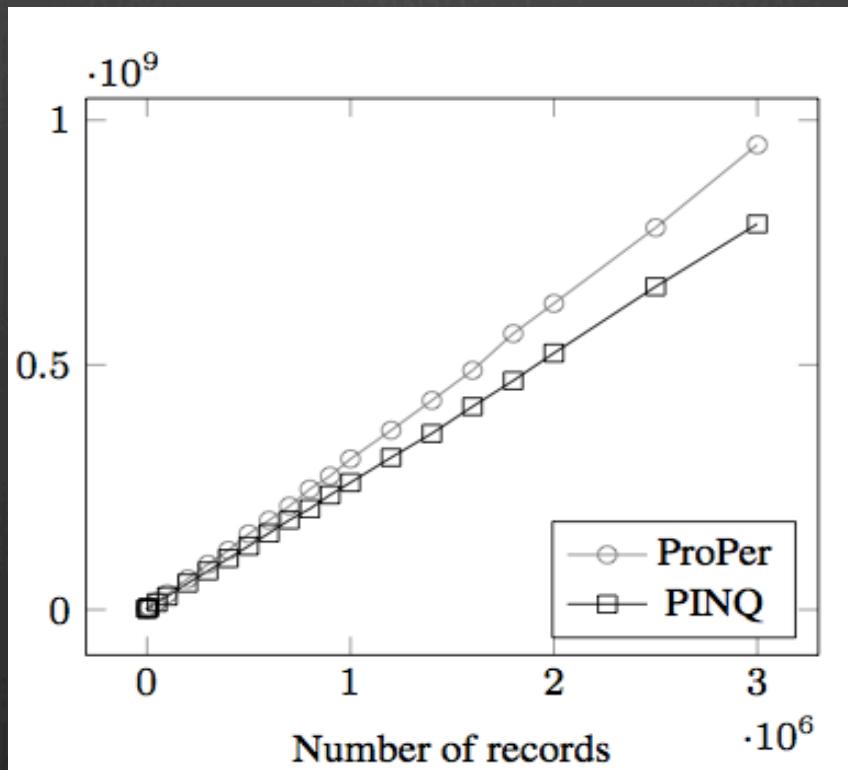
Popularity
A small icon showing two people, one red and one blue, with arrows pointing towards each other.

prov·e·nance noun \präv-nan(t)s, 'prä-vän(t)s\

: the origin or source of something

Provenance for Personalised Differential Privacy

Our implementation, ProPer, is based on (and subsumes) PINQ with small overhead



ProPer in Action

	ID	Age
1	Mary	24
1	Bob	29
1	Harry	17

Initial budgets
associated with the
original data

ProPer in Action

ID	Age
1	Mary 24
1	Bob 29
1	Harry 17

SELECT age
WHERE age \geq 18



Transformation

ProPer in Action

1
1
1

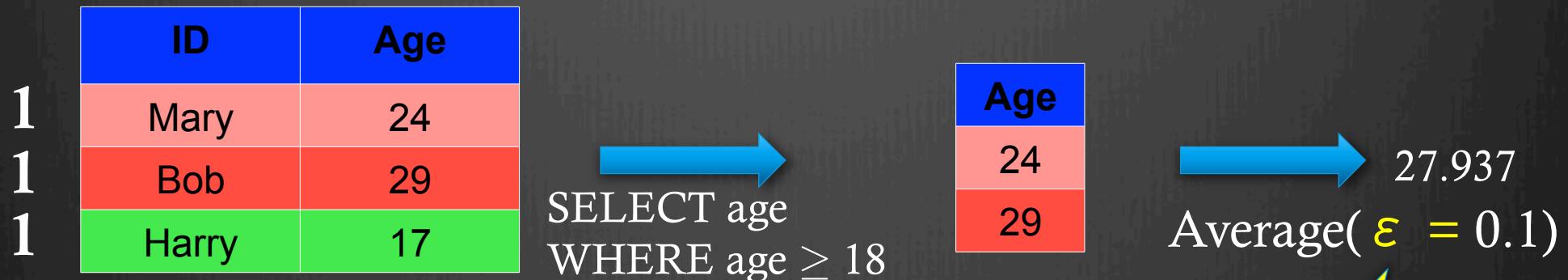
ID	Age
Mary	24
Bob	29
Harry	17

SELECT age
WHERE age \geq 18

Age
24
29

Table plus the
provenance of each
record

ProPer in Action



“Primitive”
DP-query

ProPer in Action

	ID	Age
0.9	Mary	24
0.9	Bob	29
1	Harry	17

SELECT age
WHERE age ≥ 18

Age
24
29

27.937
Average($\epsilon = 0.1$)

Update budgets

The Catch

PINQ

deny the query (throw exception)

OK because the budget is not private

ProPer

Not OK! Budget *is* private



Solution

1. Silently drop the records from the query which would otherwise get negative budget
 - Not obvious that this is privacy preserving
 - It isn't , in general
2. Restrict to *unary union-preserving transformations* (e.g. *map & filter*)
 - *Small change to dataset implies only small change to set of over-budget records*

Conclusions

- Introduced Personalised Differential Privacy
 - more fine-grained budgeting
 - capable of handling interactive queries over data arriving over time
- ProPer provenance-based tracking
 - Implementation subsuming PINQ with small overhead
 - Formal model & proof of correctness (PDP)

End



ID	Age
1	Mary 24
1	Bob 29
1	Harry 17

SELECT age
WHERE age \geq 18

Age
24
29

Average($\epsilon = 0.1$)



Further Work

- **Permissiveness:** Prove more permissive than PINQ
 - requires formal model of PINQ
- **Utility:** method degenerates to noise; analyst may be unaware
 - Track utility based on analysts prior knowledge

Related Work

- See paper
- Don't see the paper:
[Xiao & Tao, SIGMOD'06] Personalised version of k-anonymity
- [Alaggan, Gambs, Kermarrec TPDP 2015]
Heterogeneous differential privacy
- [Jorgensen, Yu, and Cormode, ICDE 2015]
Conservative or liberal? personalized differential privacy.

Related Work

Not cited in the paper: