Differential Privacy: An Economic Method for Choosing Epsilon

Justin Hsu\textsuperscript{1}  Marco Gaboardi\textsuperscript{2}
Andreas Haeberlen\textsuperscript{1}  Sanjeev Khanna\textsuperscript{1}  Arjun Narayan\textsuperscript{1}
Benjamin C. Pierce\textsuperscript{1}  Aaron Roth\textsuperscript{1}

\textsuperscript{1}University of Pennsylvania
\textsuperscript{2}University of Dundee

July 22, 2014
Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier 12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using. The data was anonymized by removing personal details and replacing names with random numbers, to protect the privacy of the recommenders.

Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas at Austin, de-anonymized some of the Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.

Their research (.pdf) illustrates some inherent security problems with anonymous data, but first it's important to explain what they did and did not do.

They did not reverse the anonymity of the entire Netflix dataset. What they did was reverse the
Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier

Last year, Netflix published a data set of about 100 million user ratings to encourage people to come up with ways to improve its algorithm. The data was anonymized to protect the privacy of its users.

Arvind Narayanan and co-workers at Stanford analyzed some of this data in their Internet Movie Database and found that it was possible to identify individual users even though their data had been anonymized.

Their research (.pdf) is important to explain why it is so hard to keep data anonymous, and it raises questions about the effectiveness of traditional privacy techniques.

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "n umb fingers" to "60 single men" to "dog that urinates on everything."
Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier

Last year, Netflix published data that was anonymized to protect the privacy of its customers. Arvind Narayanan and his colleagues released some of that data in their paper "The Internet Movie Database is not so anonymous."

Their research is important to explain why we need better protections for user privacy. They did not reveal the identity of people on Netflix, but they showed that it is possible to infer private information about users from anonymized data.

A Practical Attack to De-Anonymize Social Network Users

Gilbert Wondracek
Technical University Vienna,
Austria
{gilbert,tho}@seclab.tuwien.ac.at

Thorsten Holz

Engin Kirda
Institute Eurecom,
Sophia Antipolis
kirda@eurecom.fr

Christopher Kruegel
University of California,
Santa Barbara
chris@cs.ucsb.edu

Abstract—Social networking sites such as Facebook, LinkedIn, and Xing have been reporting exponential growth rates and have millions of registered users.

Interesting for attackers. Although social networking sites employ mechanisms to protect the privacy of their users, there is always the risk that an attacker can correlate data or

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

Eds of period on gers” to “60 single men” to
Problem: Privacy!

Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier

Last year, Netflix published a data set that was supposed to protect the privacy of individuals.

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher’s anonymity, but it was not much of a shield.

A Practical Attack to De-Anonymize Social Network Users

Gilbert Wondracek
Thorsten Holz
Technical University Vienna, Austria
{gilbert.tho}@seclab.tuwien.ac.at

Engin Kirda
Sophia Antipolis Institute Eurecom,

Christopher Kruegel
University of California, Santa Barbara

Abstract—Social networking sites such as Facebook, LinkedIn, and Xing have been reporting exponential growth rates and have millions of registered users.

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques and sparsity. Each record contains many attributes (i.e., columns in a database schema), which can be viewed as dimensions. Sparsity means that for the average record, there are no “similar” records in the multi-dimensional space defined by the attributes. This sparsity is empirically well-established [7, 4, 19] and related to the “fat
Differential privacy?

History

- Notion of privacy by Dwork, McSherry, Nissim, Smith
- Many algorithms satisfying differential privacy now known
History

- Notion of privacy by Dwork, McSherry, Nissim, Smith
- Many algorithms satisfying differential privacy now known

Some key features

- Rigorous: differential privacy must be formally proved
- Randomized: property of a probabilistic algorithm
- Quantitative: numeric measure of “privacy loss”
Differential Privacy

D

| Alice | Bob | Chris | Donna | Ernie |

Algorithm

Pr [r]

In pictures
Differential Privacy

D

Alice  Bob  Xavier  Donna  Ernie

Algorithm

Pr [r]
The setting

- Database: multiset of records (one per individual)
- Neighboring databases $D, D'$: databases differing in one record
- Randomized algorithm $M$ mapping database to outputs $\mathcal{R}$

**Definition**

Let $\varepsilon > 0$ be fixed. $M$ is $\varepsilon$-differentially private if for all neighboring databases $D, D'$ and sets of outputs $S \subseteq \mathcal{R}$,

$$\Pr[M(D) \in S] \leq e^\varepsilon \cdot \Pr[M(D') \in S].$$
The setting

- Database: multiset of records (one per individual)
- Neighboring databases $D, D'$: databases differing in one record
- Randomized algorithm $M$ mapping database to outputs $\mathcal{R}$

Definition

Let $\varepsilon > 0$ be fixed. $M$ is $\varepsilon$-differentially private if for all neighboring databases $D, D'$ and sets of outputs $S \subseteq \mathcal{R}$,

$$\Pr[M(D) \in S] \leq e^\varepsilon \cdot \Pr[M(D') \in S].$$
But what about ε?
The challenge: How to set $\varepsilon$?

The equation

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S].$$

Why do we need to set $\varepsilon$?

• Many private algorithms work for a range of $\varepsilon$, but performance highly dependent on particular choice
• Experimental evaluations of private algorithms
• Real-world uses of private algorithms
The challenge: How to set $\varepsilon$? 

The equation

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S].$$

Why do we need to set $\varepsilon$?

• Many private algorithms work for a range of $\varepsilon$, but performance highly dependent on particular choice
• Experimental evaluations of private algorithms
• Real-world uses of private algorithms
The challenge: How to set $\varepsilon$?

The equation

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S].$$

Why do we need to set $\varepsilon$?

- Many private algorithms work for a range of $\varepsilon$, but performance highly dependent on particular choice
- Experimental evaluations of private algorithms
- Real-world uses of private algorithms
Theorists say...

- Set $\varepsilon$ to be small constant, like 2 or 3
- Proper setting of $\varepsilon$ depends on society
Theorists say...

- Set $\varepsilon$ to be a small constant, like 2 or 3
- Proper setting of $\varepsilon$ depends on society

Experimentalists say...

- Try a range of values
- Literature: $\varepsilon = 0.01$ to 100
An easy question?

Theorists say...

- Set $\varepsilon$ to be a small constant, like 2 or 3
- Proper setting of $\varepsilon$ depends on society

Experimentalists say...

- Try a range of values
- Literature: $\varepsilon = 0.01$ to 100

$e^{\varepsilon} \sim 1.01$
An easy question?

Theorists say...

• Set $\varepsilon$ to be small constant, like 2 or 3
• Proper setting of $\varepsilon$ depends on society

Experimentalists say...

• Try a range of values
• Literature: $\varepsilon = 0.01$ to 100

$e^\varepsilon \sim 1.01$
$e^\varepsilon \sim 2.69 \cdot 10^{48}$
Think about costs rather than privacy

- $\varepsilon$ measures privacy, too abstract
- Monetary costs: more concrete way to measure privacy
Think about costs rather than privacy

- $\varepsilon$ measures privacy, too abstract
- Monetary costs: more concrete way to measure privacy

Add more parameters! (?)

- Break $\varepsilon$ down into more manageable parameters
- More parameters, but more concrete
- Set $\varepsilon$ as function of new parameters
Model the central tradeoff

- Stronger privacy for smaller $\varepsilon$, weaker privacy for larger $\varepsilon$
- Better accuracy for larger $\varepsilon$, worse accuracy for smaller $\varepsilon$
Model the central tradeoff

- Stronger privacy for smaller $\varepsilon$, weaker privacy for larger $\varepsilon$
- Better accuracy for larger $\varepsilon$, worse accuracy for smaller $\varepsilon$

Introduce parameters for two parties

- **Individual**: concerned about privacy
- **Analyst**: concerned about accuracy
Model the central tradeoff

- Stronger privacy for smaller $\varepsilon$, weaker privacy for larger $\varepsilon$
- Better accuracy for larger $\varepsilon$, worse accuracy for smaller $\varepsilon$

Introduce parameters for two parties

- **Individual**: concerned about privacy
- **Analyst**: concerned about accuracy

Combine the parties

- **Balance** accuracy against privacy guarantee
Model the central tradeoff

- Stronger privacy for smaller $\epsilon$, weaker privacy for larger $\epsilon$
- Better accuracy for larger $\epsilon$, worse accuracy for smaller $\epsilon$

Introduce parameters for two parties

- **Individual**: concerned about privacy
- **Analyst**: concerned about accuracy

Combine the parties

- **Balance** accuracy against privacy guarantee
What does $\varepsilon$ mean for privacy?

Certainly, it is the case — Excuse me, but I beg to differ... Fools!

\[ \frac{\varepsilon}{3} \quad \frac{\varepsilon}{3} \quad \frac{\varepsilon}{3} \]

($\varepsilon$-over-3 argument)

2008
Participation

- Private algorithm $M$ is a study
- Bob the individual has choice to participate in the study
- Study will happen regardless of Bob’s choice
Interpreting \( \varepsilon \)

**Participation**
- Private algorithm \( M \) is a *study*
- Bob the individual has choice to participate in the study
- Study will happen regardless of Bob’s choice

**Bad events**
- Set of real-world *bad events* \( \mathcal{O} \)
- Bob wants to avoid these events
Thought experiment: two possible worlds

- Identical, except Bob participates in first world and not in the second world
- Rest of database, all public information is identical
- All differences in two worlds due to the output of the study
- Every output $r \in \mathcal{R}$ leads to an event in $\mathcal{O}$ or not
For all sets of outputs $S$...

$$\Pr[M(D) \in S] \leq e^\epsilon \cdot \Pr[M(D') \in S].$$
For all sets of outputs $S$...

$$\Pr[M(D) \in S] \leq e^\varepsilon \cdot \Pr[M(D') \in S].$$

Participate

Don't participate

Bad events interpretation of $\varepsilon$
For all sets of outputs $S$...

$$\Pr[M(D) \in S] \leq e^\varepsilon \cdot \Pr[M(D') \in S].$$
For all sets of outputs $S$...

$$\Pr[M(D) \in S] \leq e^\varepsilon \cdot \Pr[M(D') \in S].$$

**Bad events interpretation of $\varepsilon$**

- Let $S$ be set of outputs leading to events in $\mathcal{O}$
- Bob participating increases probability of bad event by at most $e^\varepsilon$ factor
Bad events not equally bad

- Cost function on bad events $f : \mathcal{O} \rightarrow \mathbb{R}^+$ (non-negative)
- Insurance premiums, embarrassment, etc.
Introducing cost

Bad events not equally bad

- Cost function on bad events $f: \mathcal{O} \to \mathbb{R}^+$ (non-negative)
- Insurance premiums, embarrassment, etc.

Our model

Pay participants for their cost
Marginal increase in cost

- Someone (society?) has decided the study is worth running
- Non-participants may feel cost, but are not paid
- Only pay participants for \textit{increase} in expected cost
How much to pay?

Marginal increase in cost

- Someone (society?) has decided the study is worth running
- Non-participants may feel cost, but are not paid
- Only pay participants for increase in expected cost

The cost of participation

- Can show: under $\varepsilon$-differential privacy, expected cost increase is at most $e^\varepsilon$ factor when participating
- Non-participants: expected cost $P$
- Participants: expected cost at most $e^\varepsilon P$
- Compensate participants: $e^\varepsilon P - P$
Individuals

• have an expected cost $P$ if they do not participate, determined by their cost function;

• can choose to participate in an $\varepsilon$-private study for fixed $\varepsilon$ in exchange for fixed monetary payment;

• participate if payment is larger than their increase in expected cost for participating: $e^\varepsilon P - P$. 
Individuals

• have an expected cost $P$ if they do not participate, determined by their cost function;
• can choose to participate in an $\varepsilon$-private study for fixed $\varepsilon$ in exchange for fixed monetary payment;
• participate if payment is larger than their increase in expected cost for participating: $e^\varepsilon P - P$.

How to set $P$?

• Depends on people’s perception of privacy costs
• Derive empirically, surveys
Summing up: the individual model

Individuals

• have an expected cost $P$ if they do not participate, determined by their cost function;
• can choose to participate in an $\varepsilon$-private study for fixed $\varepsilon$ in exchange for fixed monetary payment;
• participate if payment is larger than their increase in expected cost for participating: $e^\varepsilon P - P$.

How to set $P$?

• Depends on people’s perception of privacy costs
• Derive empirically, surveys
Model the central tradeoff

- Stronger privacy for smaller $\varepsilon$, weaker privacy for larger $\varepsilon$
- Better accuracy for larger $\varepsilon$, worse accuracy for smaller $\varepsilon$

Introduce parameters for two parties

- **Individual**: concerned about privacy
- **Analyst**: concerned about accuracy

Combine the parties

- **Balance** accuracy against privacy guarantee
The plan today

Model the central tradeoff

- Stronger privacy for smaller $\varepsilon$, weaker privacy for larger $\varepsilon$
- Better accuracy for larger $\varepsilon$, worse accuracy for smaller $\varepsilon$

Introduce parameters for two parties

- **Individual**: concerned about privacy
- **Analyst**: concerned about accuracy

Combine the parties

- **Balance** accuracy against privacy guarantee
Why not just take $\varepsilon$ small?
Accuracy?

- Study is run to learn some information; want useful results
- Setting $\varepsilon$ small will be very private, but very inaccurate (?)
Accuracy?

- Study is run to learn some information; want useful results
- Setting $\varepsilon$ small will be very private, but very inaccurate (?)

Another parameter: the study size $N$

- Natural parameter of the study, measures amount of data
- Typical studies: accuracy improves as $N$ increases
Introducing the analyst

Alice the analyst

- Has a private study $M$, works for range of $\varepsilon$ and study size $N$
- Wants to set these two parameters
- Has numeric measure of accuracy for this study
- Wants to achieve set level of accuracy
Alice the analyst

- Has a private study $M$, works for range of $\varepsilon$ and study size $N$
- Wants to set these two parameters
- Has numeric **measure of accuracy** for this study
- Wants to achieve set **level of accuracy**
What is accuracy?

Measure of accuracy

- Real number, depends on the study $M$, parameters $\varepsilon$ and $N$
- Could be defined as:
  - Distance from true answer
  - Probability of exceeding error
  - Number of mistakes
  - ...  

Level of accuracy

- Real number, maximum allowable accuracy
- Captures Alice’s requirement for the study
Summing up: The analyst model

The analyst

- has an $\varepsilon$-private study $M$;
- has a numeric measure of accuracy $A_M(\varepsilon, N) : \mathbb{R}$;
- has a numeric accuracy level $T : \mathbb{R}$;
- wants $A_M(\varepsilon, N) \leq T$. 

How to set $A_M$?
- Theoretical accuracy guarantee for $M$ from literature
- Empirical trials: measure accuracy of $M$ on test data

How to set $T$?
- Ask the analyst what accuracy is needed
The analyst

- has an $\varepsilon$-private study $M$;
- has a numeric measure of accuracy $A_M(\varepsilon, N) : \mathbb{R}$;
- has a numeric accuracy level $T : \mathbb{R}$;
- wants $A_M(\varepsilon, N) \leq T$.

How to set $A_M$?

- Theoretical accuracy guarantee for $M$ from literature
- Empirical trials: measure accuracy of $M$ on test data
Summing up: The analyst model

The analyst

• has an $\varepsilon$-private study $M$;
• has a numeric measure of accuracy $A_M(\varepsilon, N) : \mathbb{R}$;
• has a numeric accuracy level $T : \mathbb{R}$;
• wants $A_M(\varepsilon, N) \leq T$.

How to set $A_M$?

• Theoretical accuracy guarantee for $M$ from literature
• Empirical trials: measure accuracy of $M$ on test data

How to set $T$?

• Ask the analyst what accuracy is needed
Model the central tradeoff

- Stronger privacy for smaller $\varepsilon$, weaker privacy for larger $\varepsilon$
- Better accuracy for larger $\varepsilon$, worse accuracy for smaller $\varepsilon$

Introduce parameters for two parties

- Individual: concerned about privacy
- Analyst: concerned about accuracy

Combine the parties

- Balance accuracy against privacy guarantee
The plan today

Model the central tradeoff

- Stronger privacy for smaller $\varepsilon$, weaker privacy for larger $\varepsilon$
- Better accuracy for larger $\varepsilon$, worse accuracy for smaller $\varepsilon$

Introduce parameters for two parties

- **Individual**: concerned about privacy
- **Analyst**: concerned about accuracy

Combine the parties

- **Balance** accuracy against privacy guarantee
Finally, how to set $\varepsilon$?
Combining the two parties

Budget

- Analyst has budget $B$ (charge it to the grant!)
- Pays sufficient compensation to all $N$ individuals

The goal: find $\varepsilon$ and $N$ such that

- Study is accurate enough
- Analyst has enough budget to pay all individuals
System of constraints

1. Accuracy constraint:
   
   \[ A_M(\varepsilon, N) \leq T \]

2. Budget constraint:
   
   \[ (e^\varepsilon P - P) \cdot N \leq B \]
Setting $\varepsilon$

System of constraints

1. Accuracy constraint:

$$A_M(\varepsilon, N) \leq T$$

2. Budget constraint:

$$(e^\varepsilon P - P) \cdot N \leq B$$

Variables

- Both sides want to find mutually agreeable setting of $\varepsilon$
- Analyst also wants to find appropriate study size $N$
- Study feasible $\iff$ constraints satisfiable
Setting $\varepsilon$

System of constraints

1. Accuracy constraint:

$$A_M(\varepsilon, N) \leq T$$

2. Budget constraint:

$$(e^\varepsilon P - P) \cdot N \leq B$$

Variables

- Both sides want to find mutually agreeable setting of $\varepsilon$
- Analyst also wants to find appropriate study size $N$
- Study feasible $\iff$ constraints satisfiable

Set $\varepsilon$ (and $N$) to satisfy constraints
Case studies: See paper!
Extending the model

In the paper

- Handle \((\varepsilon, \delta)\)-privacy
- Add other constraints: limit size of study
- Rule out values of \(\varepsilon\) that aren’t “intuitively” private
Extending the model

In the paper

- Handle $(\varepsilon, \delta)$-privacy
- Add other constraints: limit size of study
- Rule out values of $\varepsilon$ that aren’t “intuitively” private

Further refinements?

- Handle collusion among participants
- Model large $\varepsilon$ regime better
Take-away points

- Parameter $\varepsilon$ is too abstract
- Use economic cost as a measure of privacy
- Use more concrete parameters: costs, budgets, accuracy, etc.
Take-away points

- Parameter $\varepsilon$ is too abstract
- Use economic cost as a measure of privacy
- Use more concrete parameters: costs, budgets, accuracy, etc.

Going forward

- More empirical research: How do people perceive costs?
- Practical attacks on $\varepsilon$-differential privacy? For what $\varepsilon$? For what algorithms?
Differential Privacy: An Economic Method for Choosing Epsilon

Justin Hsu\textsuperscript{1} Marco Gaboardi\textsuperscript{2}
Andreas Haeberlen\textsuperscript{1} Sanjeev Khanna\textsuperscript{1} Arjun Narayan\textsuperscript{1}
Benjamin C. Pierce\textsuperscript{1} Aaron Roth\textsuperscript{1}

\textsuperscript{1}University of Pennsylvania
\textsuperscript{2}University of Dundee

July 22, 2014
Which events are considered?

Key assumption: participation decision

- Bob’s choice only visible via the output of the study
- Arbitrary side information may be public, as long as it is the same whether Bob participates or not
- Crucial for differential privacy to give a meaningful guarantee!
Which events are considered?

Key assumption: participation decision

- Bob’s choice only visible via the output of the study
- Arbitrary side information may be public, as long as it is the same whether Bob participates or not
- Crucial for differential privacy to give a meaningful guarantee!

No “side-channels”
Which events are considered?

Key assumption: participation decision

- Bob’s choice only visible via the output of the study
- Arbitrary side information may be public, as long as it is the same whether Bob participates or not
- Crucial for differential privacy to give a meaningful guarantee!

Example: non-protected event

- Someone...
  - monitors Bob’s bank account and sees payment for study;
  - or sees Bob participating in the study;
- ...then uses output of study to break Bob’s privacy
Individuals With Different Costs?

- Individuals may have different cost functions $f$
- But cost function may be private, correlated with private data
- Not clear how to compensate them differently, so pay each individual the same amount $C$

Sampling Bias

- Setting $C$ too low can skew database towards people who don’t have very high cost
- Ideal: $C$ is the maximum increase in expected cost $P$
Setting: Bob the Individual

- Insurance companies don’t know Bob smokes
- Bob is worried about his insurance premium increasing
Case Study: Estimating A Mean

Setting: Bob the Individual

- Insurance companies don’t know Bob smokes
- Bob is worried about his insurance premium increasing

Setting: Alice the Analyst

- Alice conducting a study on medical records
- Goal: estimate the fraction of the patients who smoke
- Must work under $\varepsilon$-differential privacy
Adding Noise

- Want to compute fraction $x$, but privately
- Say $x$ can differ by $\Delta$ on neighboring databases
- Draw noise $\nu$ from the Laplace distribution with scale $\Delta/\varepsilon$
- Releasing $x + \nu$ is $\varepsilon$-differentially private
Adding Noise

- Want to compute fraction $x$, but privately
- Say $x$ can differ by $\Delta$ on neighboring databases
- Draw noise $\nu$ from the Laplace distribution with scale $\Delta/\varepsilon$
- Releasing $x + \nu$ is $\varepsilon$-differentially private

Figure: Laplace distribution
Adding Noise

- Want to compute fraction $x$, but privately
- Say $x$ can differ by $\Delta$ on neighboring databases
- Draw noise $\nu$ from the Laplace distribution with scale $\Delta/\varepsilon$
- Releasing $x + \nu$ is $\varepsilon$-differentially private

\[ \Pr \]

$\Delta/\varepsilon$

0 $\nu$

\textbf{Figure:} Laplace distribution
Adding Noise

- Want to compute fraction $x$, but privately
- Say $x$ can differ by $\Delta$ on neighboring databases
- Draw noise $\nu$ from the Laplace distribution with scale $\Delta/\varepsilon$
- Releasing $x + \nu$ is $\varepsilon$-differentially private

**Figure**: Laplace distribution
What is the Cost of Not Participating $P$?

- Correct way to estimate parameter: conduct surveys
What is the Cost of Not Participating $P$?

- Correct way to estimate parameter: conduct surveys
- Our bad event: health insurance premium increase ($1274$)
What is the Cost of Not Participating $P$?

- Correct way to estimate parameter: conduct surveys
- Our bad event: health insurance premium increase ($1274)
- Bob estimates probability this happens even if he doesn’t participate: 5%

Expected cost of non-participation:

\[ P \cdot \$1274 = \$63.7 \]

Bob will participate if paid $63.7 \cdot (e^\varepsilon - 1)$.
What is the Cost of Not Participating $P$?

- Correct way to estimate parameter: conduct surveys
- Our bad event: health insurance premium increase ($1274)
- Bob estimates probability this happens even if he doesn’t participate: 5%
- Expected cost of non-participation: $P = 5\% \cdot 1274 = 63.7$
What is the Cost of Not Participating $P$?

- Correct way to estimate parameter: conduct surveys
- Our bad event: health insurance premium increase ($1274)
- Bob estimates probability this happens even if he doesn’t participate: 5%
- Expected cost of non-participation: $P = 5\% \cdot 1274 = 63.7$

Bob will participate if paid $63.7 \cdot (e^\varepsilon - 1)$
Measuring the Accuracy

- Alice wants fraction of smokers to within 0.05 error
- Measure of accuracy: $A_M(\varepsilon, N)$ is probability of exceeding this error, want probability to be small (at most 10% chance)
Measuring the Accuracy

- Alice wants fraction of smokers to within 0.05 error
- Measure of accuracy: $A_M(\varepsilon, N)$ is probability of exceeding this error, want probability to be small (at most 10% chance)

Dependence on Database Size

- Changing one record changes $\mu$ by at most $1/N$
- As $N$ grows, less noise needed for $\varepsilon$-privacy
The Budget Constraint

- Alice has $B = 30,000 to spend: constraint

$$63.7 \cdot (e^\varepsilon - 1) \cdot N \leq 30000$$
Applying the Model

The Budget Constraint

- Alice has $B = 30,000 to spend: constraint

  \[ 63.7 \cdot (e^\varepsilon - 1) \cdot N \leq 30000 \]

The Accuracy Constraint

- Alice wants probability of exceeding error at most 10%
- Sets $T = 0.1$ and requires $A_M(\varepsilon, N) \leq T = 0.1$
- Can be shown via statistical tools, sufficient to have

  \[ 2 \exp(-0.0002N) + \exp(-0.025N\varepsilon) \leq 0.1 \]
Applying the Model

The Budget Constraint

- Alice has $B = 30,000 to spend: constraint

\[ 63.7 \cdot (e^\varepsilon - 1) \cdot N \leq 30000 \]

The Accuracy Constraint

- Alice wants probability of exceeding error at most 10%
- Sets $T = 0.1$ and requires $A_M(\varepsilon, N) \leq T = 0.1$
- Can be shown via statistical tools, sufficient to have

\[ 2 \exp (-0.0002N) + \exp (-0.025N\varepsilon) \leq 0.1 \]

Study feasible $\iff$ constraints satisfiable
Is the Study Feasible?

Yes!

- $N = 15000$, $\varepsilon = 0.03$

- Bob is paid $1.93$

Figure: Feasible $\varepsilon$, $N_B$, for accuracy $T$ and budget $B$. 
Is the Study Feasible?

Yes!

- $N = 15000, \varepsilon = 0.03$
- Bob is paid $1.93$
Is the Study Feasible?

Yes!

- $N = 15000, \varepsilon = 0.03$
- Bob is paid $1.93

**Figure:** Feasible $\varepsilon, N$, for accuracy $T$ and budget $B$. 
Other Applications: The Cost of Privacy

Non-private Studies

- No privacy guarantee
- What if non-private studies had to pay extra for this risk?

Tradeoff

- Non-private study has better accuracy, need smaller study, but needs to pay more per person
- Private study has worse accuracy, needs bigger study, but pays less per person

Our Model

- Private study is sometimes cheaper than equivalent non-private study!