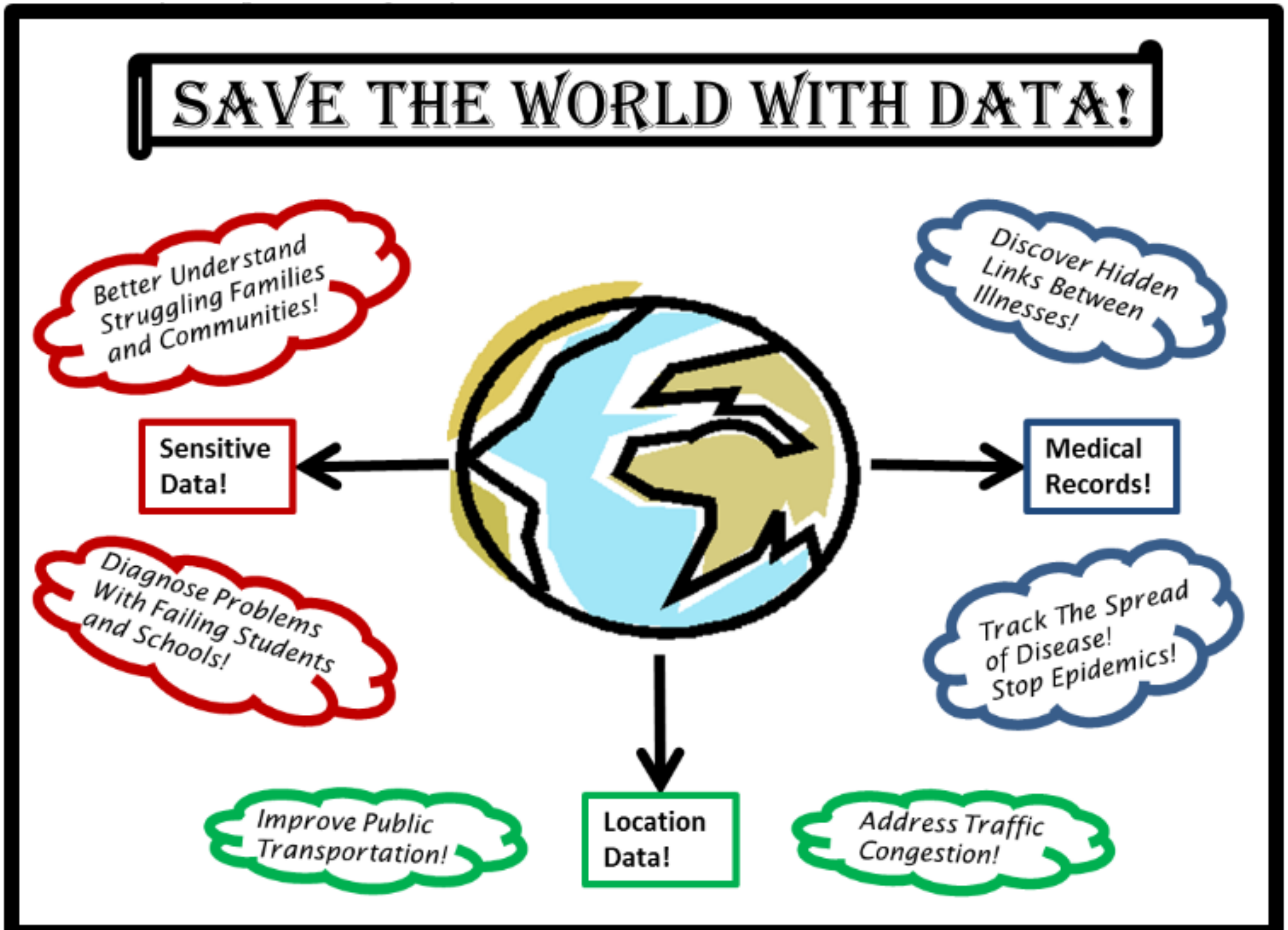
Privacy-preserving Datamining: Differential Privacy And Applications

Christine Task
PhD Candidate
Computer Science Department
Purdue University
Advisor: Chris Clifton

In The Era of Big Data…



Presentation Outline

- Definitions
- Basic Use
- Applications: Social Network Analysis
- Applications: Learning Analytics

Definitions

You're handed a survey…

1) Do you like listening to Justin Bieber?

2) How many Justin Bieber albums do you own?

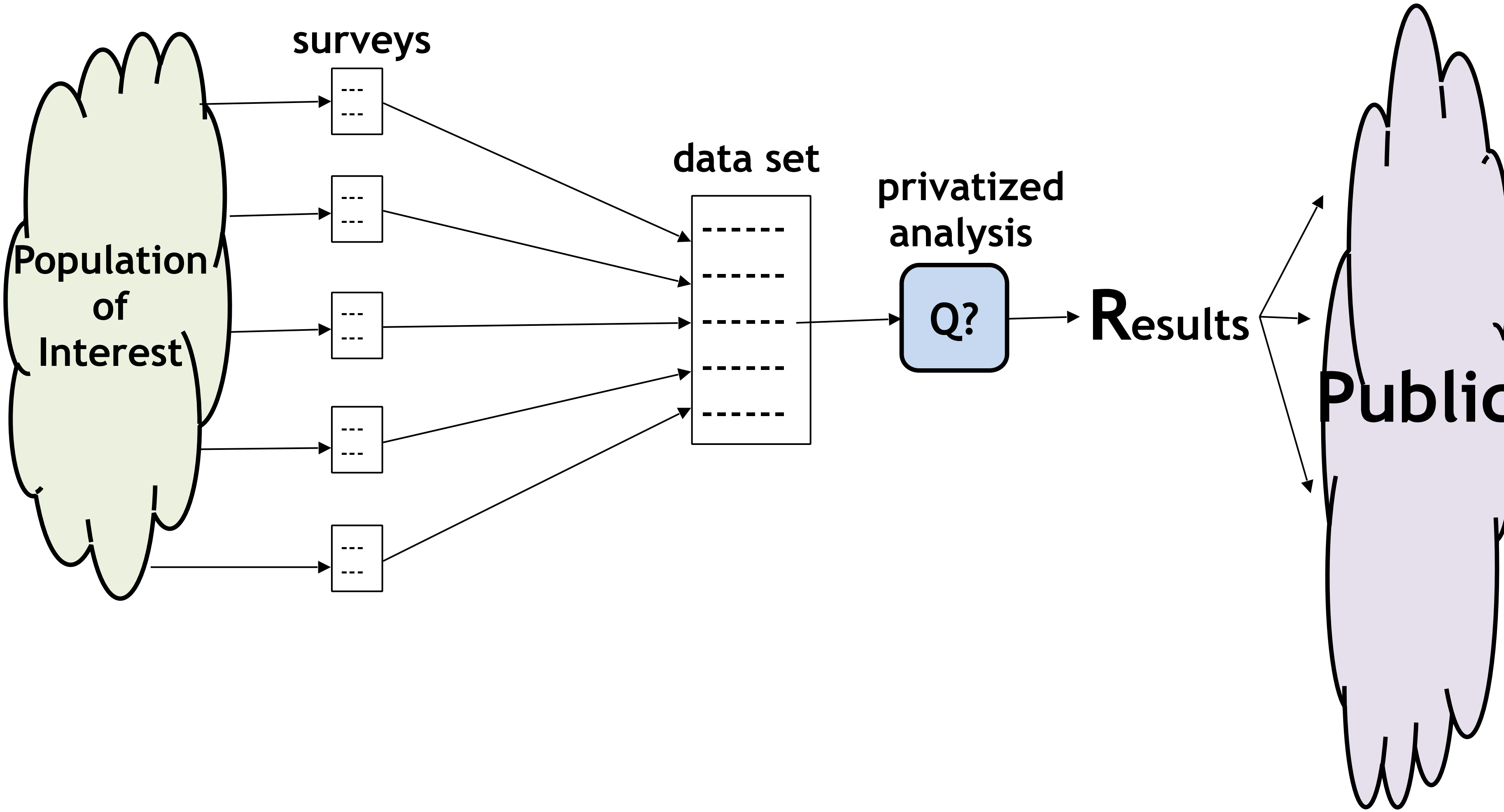
3) What is your gender?

4) What is your age?

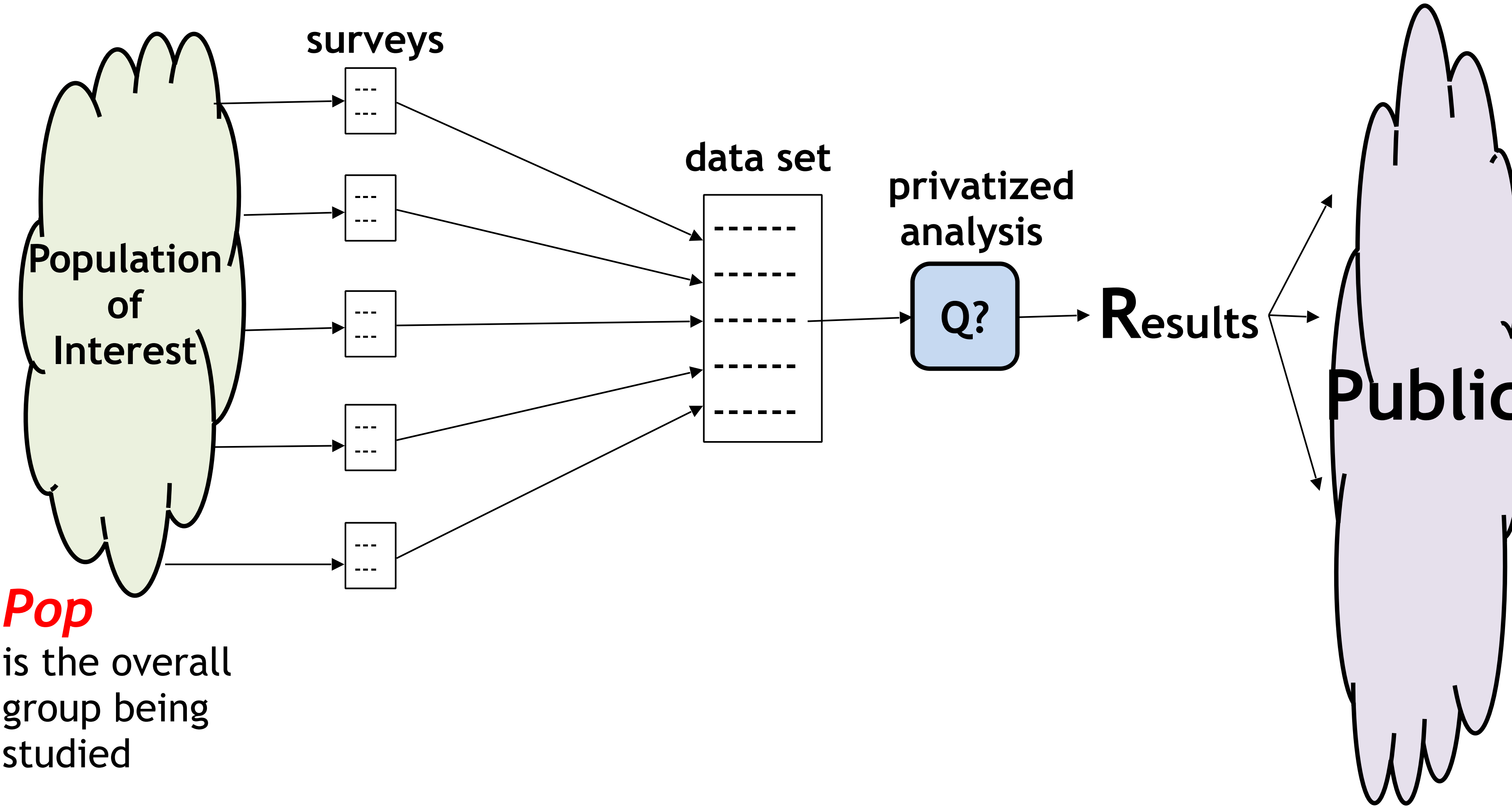
The researcher tells you the data from the surveys will be collected into a data-set, then some analysis will be done and the results released to the public. She says it's perfectly safe to submit a survey: it's anonymous and the analysis will be privatized.

What do you do?

The Notation

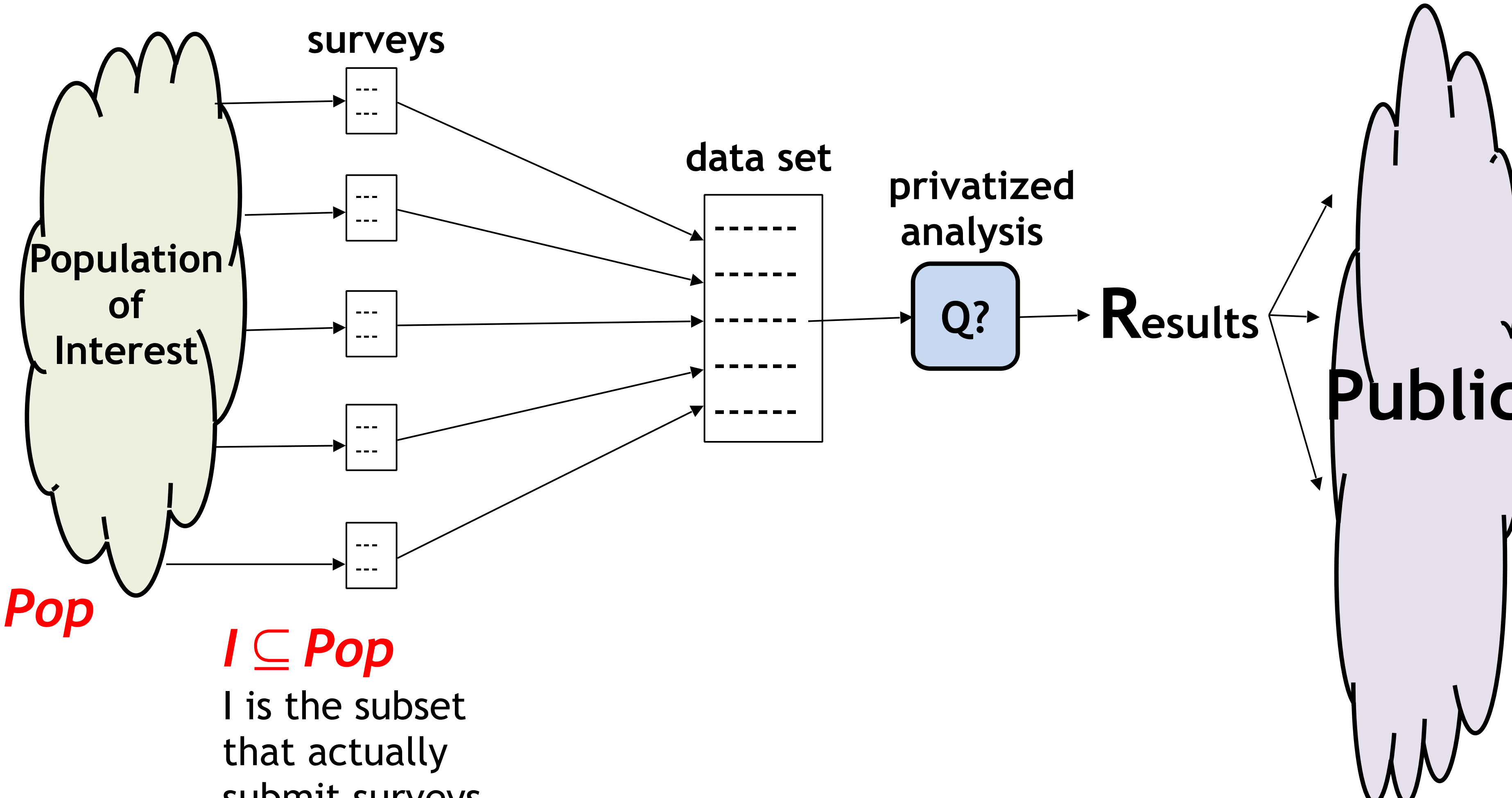


The Notation



Pop is the overall group being studied

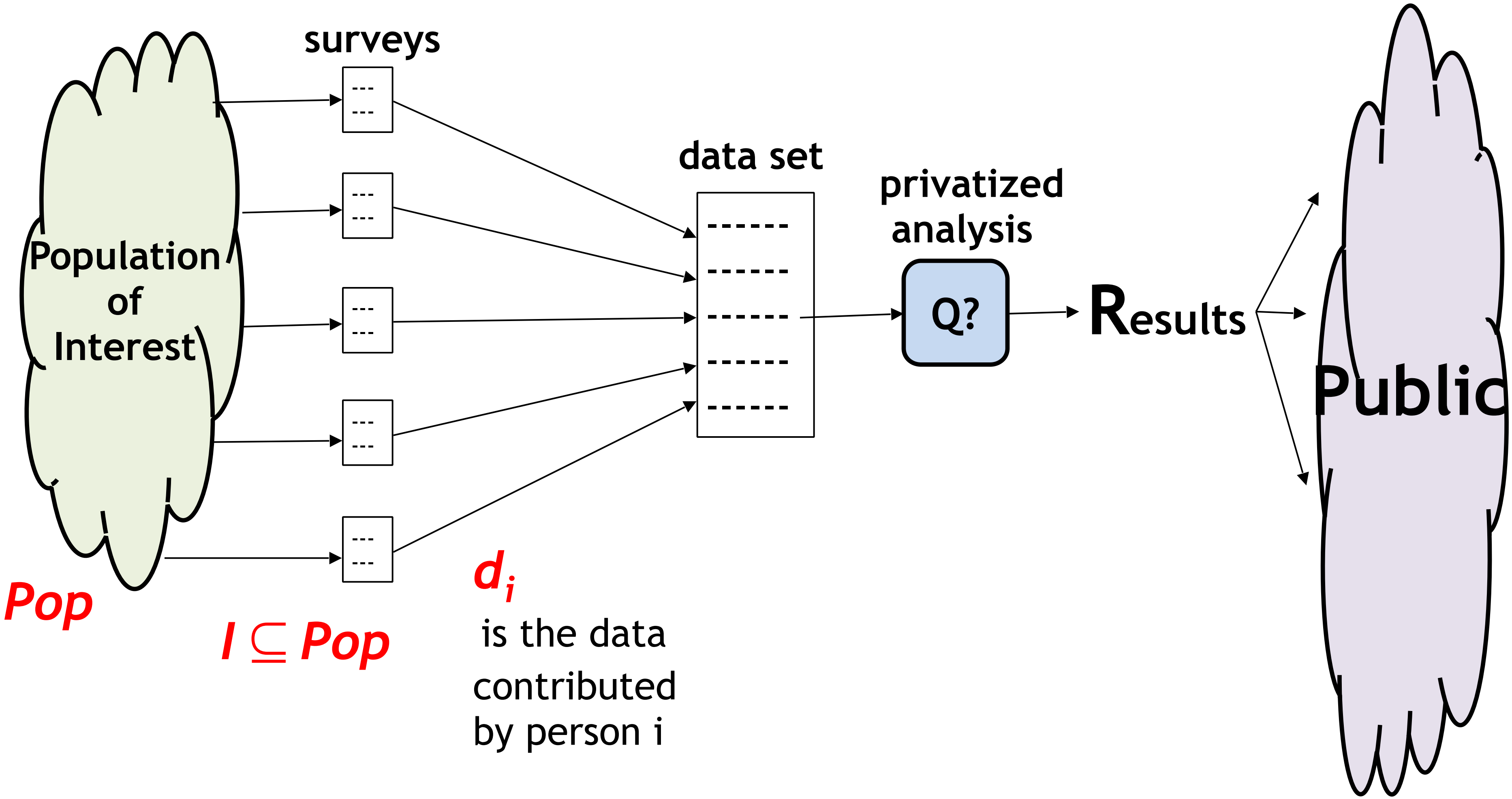
The Notation



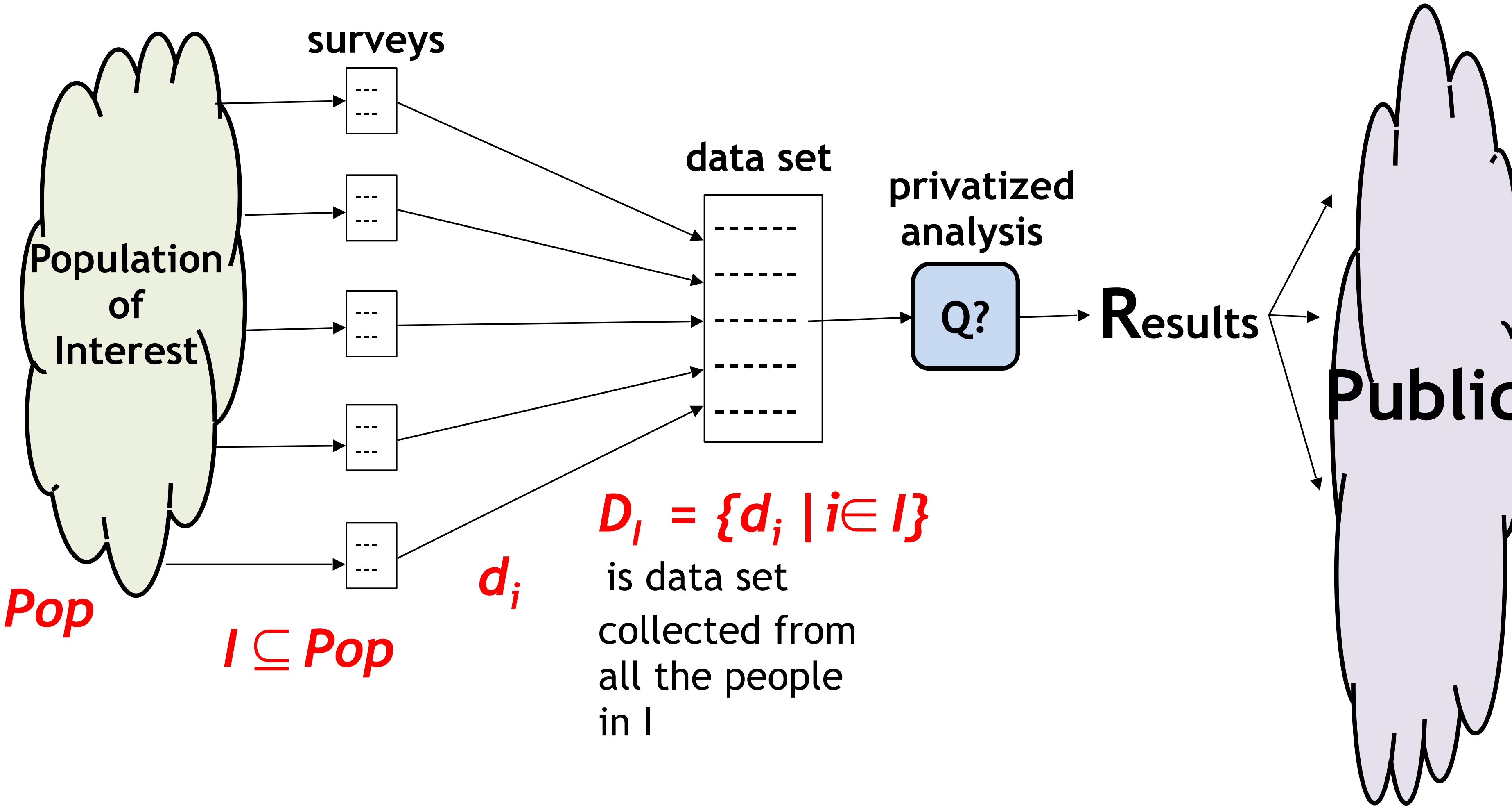
I ⊆ *Pop*

I is the subset that actually submit surveys

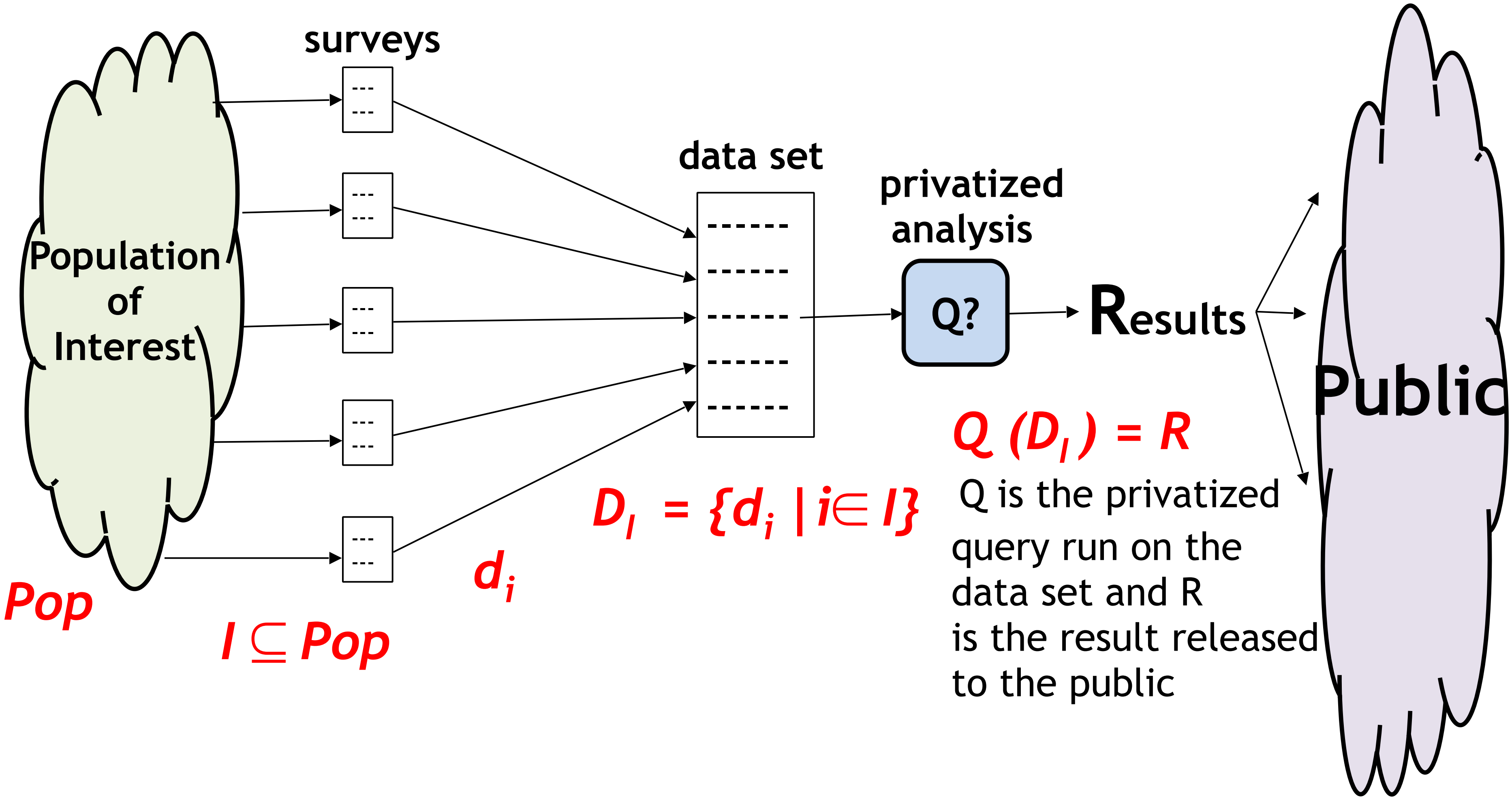
The Notation



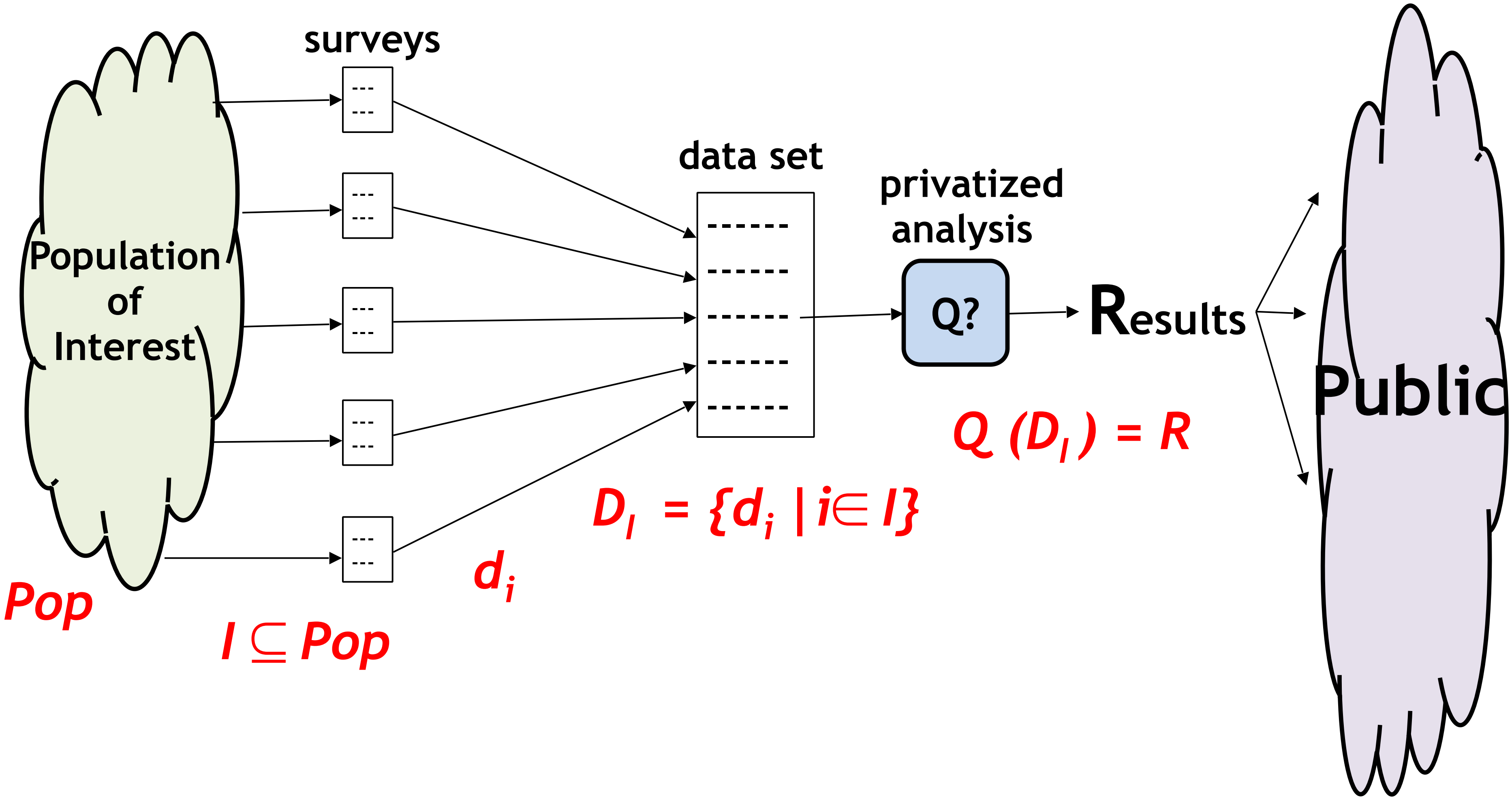
The Notation



The Notation



The Notation



What do we want?

I would feel safe submitting a survey if…

What do we want?

I would feel safe submitting a survey if…

- I knew that my answer had no impact on the released results.

- $Q(D_{(I\text{-}me)}) = Q(D_I)$

What do we want?

I would feel safe submitting a survey if…

- I knew that my answer had no impact on the released results.
- I knew that any attacker looking at the published results R couldn't learn (with high probability) any new information about me personally.

- $Q(D_{(I\text{-}me)}) = Q(D_I)$
- $\text{Prob}(\textit{secret}(\text{me}) \mid R) = \text{Prob}(\textit{secret}(\text{me}))$

Why can’t we have it?

Why can't we have it?

- If individual answers had no impact on the released results… Then the results would have no utility

- By induction,
  $Q(D_{(I\text{-}me)}) = Q(D_I) \Rightarrow$
  $Q(D_I) = Q(D_{\varnothing})$

Why can't we have it?

- If individual answers had no impact on the released results… Then the results would have no utility

- If R shows there's a strong trend in my population, then with high probability, the trend is true of me too (even if I don't submit a survey).

- By induction, $Q(D_{(I\text{-}me)}) = Q(D_I) \Rightarrow Q(D_I) = Q(D_\varnothing)$

- $\text{Prob}(\textit{secret}(\text{me}) \mid \textit{secret}(\text{Pop})) > \text{Prob}(\textit{secret}(\text{me}))$

Why can't we have it?

- Even worse, if an attacker knows a function about me that's dependent on general facts about the population:

  - **I'm twice the average age**
  - **I'm in the minority gender**

  Then releasing just those general facts gives the attacker specific information about me. (Even if I don't submit a survey!)

- $(\mathit{age}(\text{me}) = 2 * \text{mean\_age}) \wedge$
  $(\mathit{gender}(\text{me}) \neq \text{mode\_gender}) \wedge$
  $(\text{mean\_age} = 14) \wedge$
  $(\text{mode\_gender} = F) \Rightarrow$

  $(\mathit{age}(\text{me}) = 28) \wedge$
  $(\mathit{gender}(\text{me}) = M)$

One more try…

So we can’t promise that my data won’t affect the results,

One more try…

So we can’t promise that my data won’t affect the results,

And we can’t promise that an attacker won’t be able to learn new information about me from looking at the results.

One more try…

So we can’t promise that my data won’t affect the results,

And we can’t promise that an attacker won’t be able to learn new information about me from looking at the results,

So what ***can*** we do?

One more try…

I'd feel safe submitting a survey if….

When the researchers published the (privatized, noisy) result R, I knew they were:
"***just about as likely to get R for their answer whether or not I submitted my information***"
... so I might as well submit

Differential Privacy

Differential Privacy is a *Guarantee* from the researcher to the individuals in the data set:

The chance that the noisy released result will be R is nearly the same, whether or not you submit your information.

$$\boxed{\frac{Prob(Q(D_I) = R)}{Prob(Q(D_{I \pm i}) = R)} \leq A, \qquad for\ all\ I, i, R}$$

Q is the query algorithm, which includes randomized noise for privatization.

A is a value close to 1 which is chosen by the researcher. When A is much larger than 1, very little privacy is offered. If A=1, then individuals have no effect on the results and there is zero utility. Formally, we define $\boxed{A = e^{\varepsilon} \text{ for small } \varepsilon > 0,}$ which is mathematically convenient, as we'll demonstrate later.

Differential Privacy

The chance that the noisy released result will be R is nearly the same, whether or not you submit your information.

$$\frac{Prob(R \mid true\ world\ = D_I)}{Prob(R \mid true\ world = D_{I\pm i})} \leq e^{\varepsilon}, \qquad for\ all\ I, i, R\ and\ small\ \varepsilon > 0$$

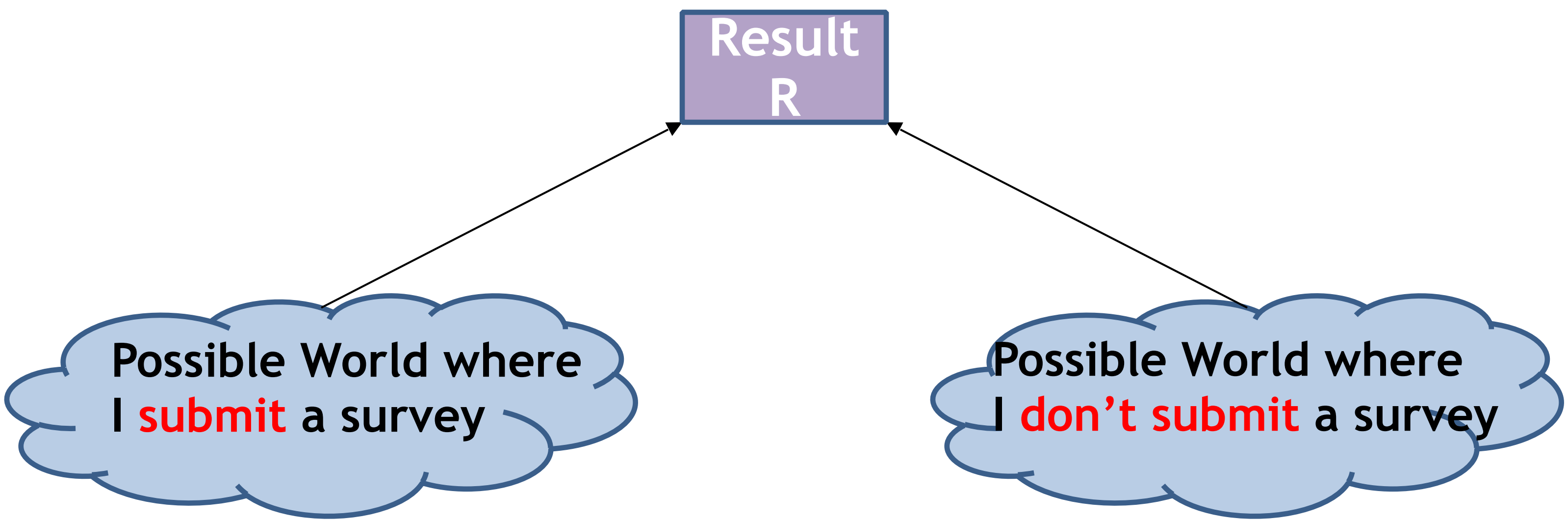
Possible World where
I submit a survey

Possible World where
I don't submit a survey

Differential Privacy

The chance that the noisy released result will be R is nearly the same, whether or not you submit your information.

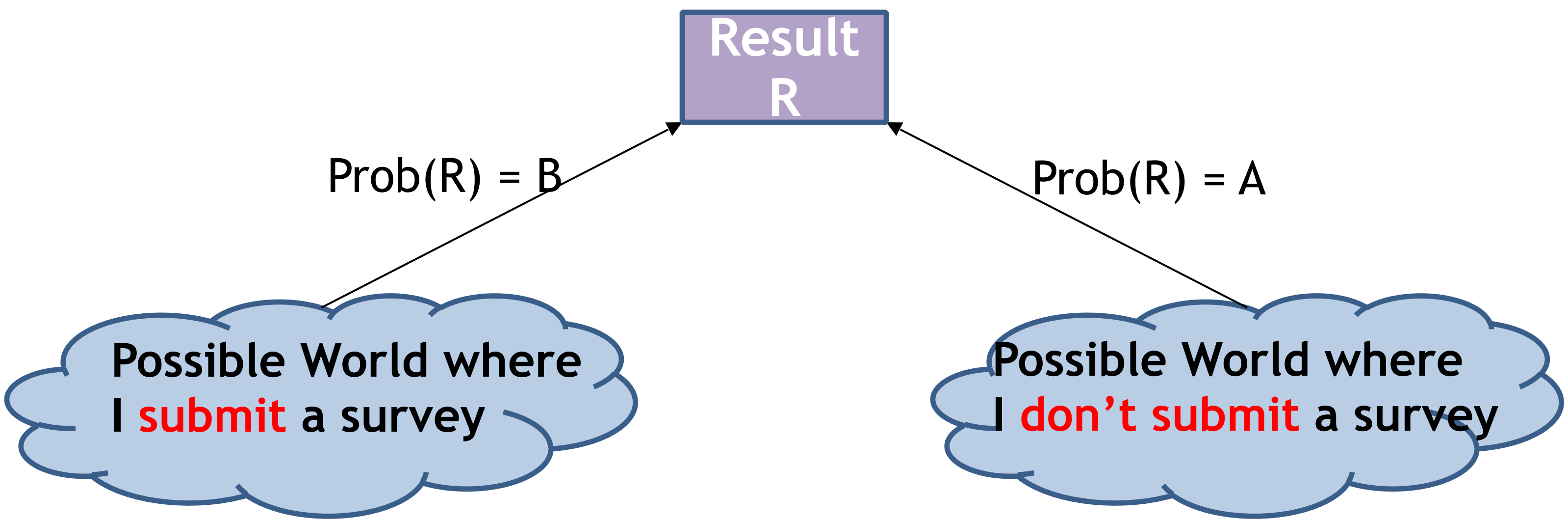
$$\frac{Prob(R \mid true\ world\ = D_I)}{Prob(R \mid true\ world = D_{I\pm i})} \leq e^{\varepsilon}, \qquad for\ all\ I, i, R\ and\ small\ \varepsilon > 0$$



Differential Privacy

The chance that the noisy released result will be R is nearly the same, whether or not you submit your information.

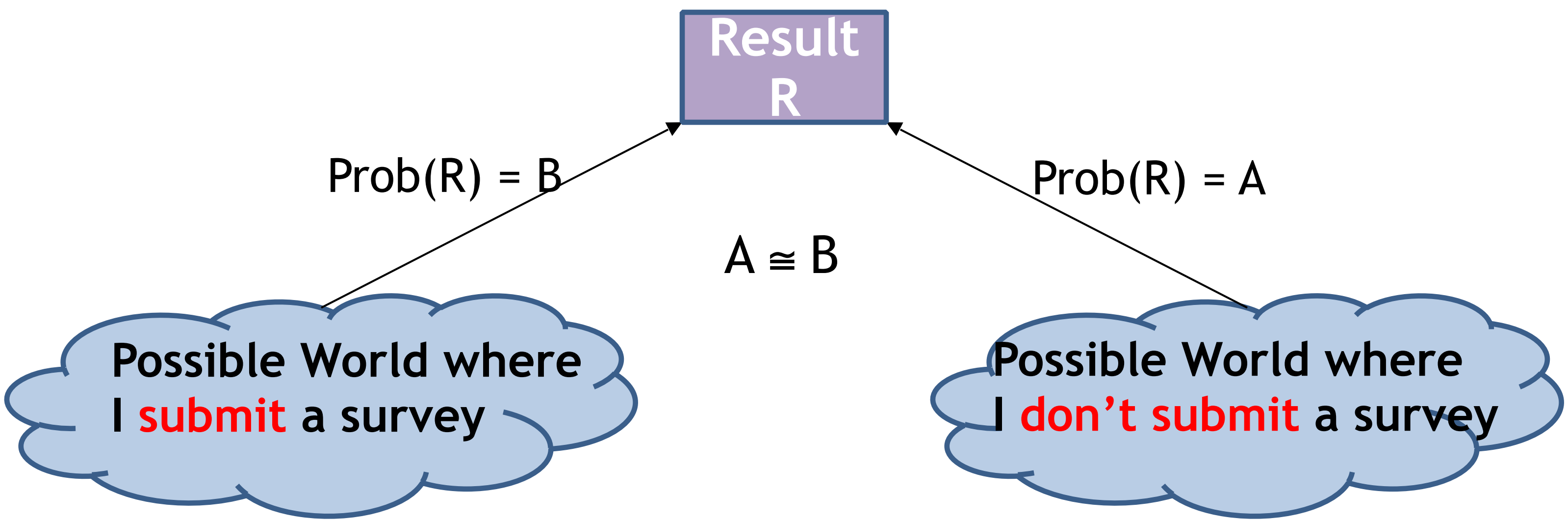
$$\frac{Prob(R \mid true\ world\ = D_I)}{Prob(R \mid true\ world = D_{I\pm i})} \leq e^{\varepsilon}, \qquad for\ all\ I, i, R\ and\ small\ \varepsilon > 0$$



Differential Privacy

The chance that the noisy released result will be R is nearly the same, whether or not you submit your information.

$$\frac{Prob(R \mid true\ world\ = D_I)}{Prob(R \mid true\ world = D_{I\pm i})} \leq e^{\varepsilon}, \qquad for\ all\ I, i, R\ and\ small\ \varepsilon > 0$$

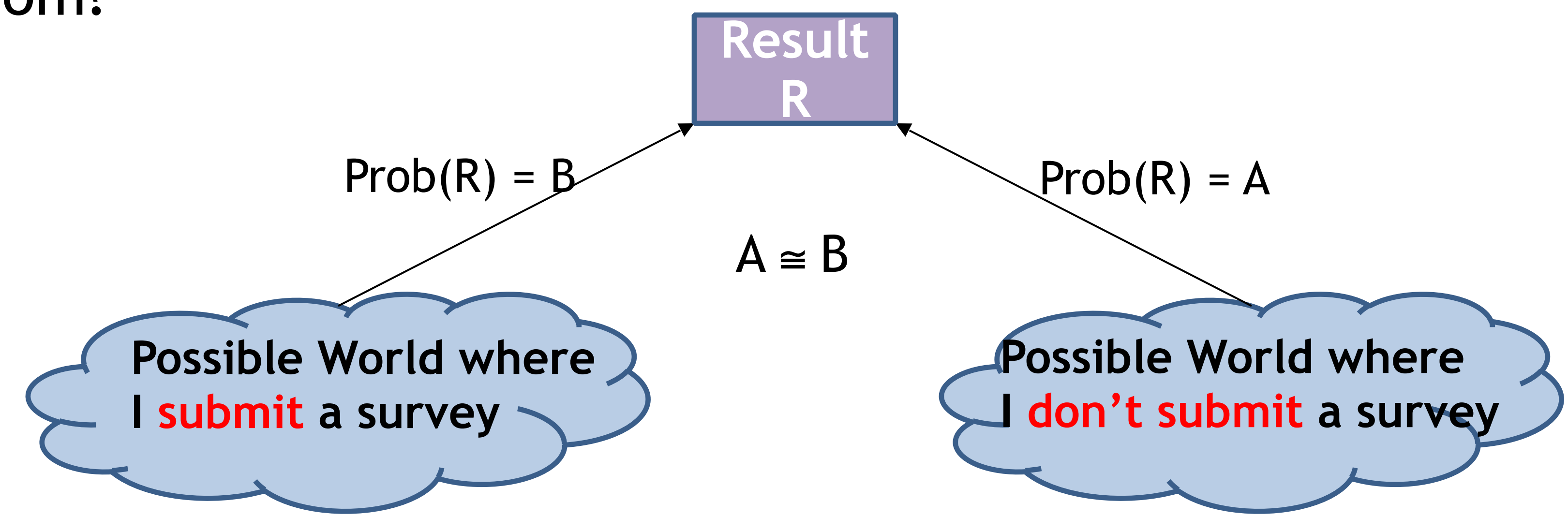


Differential Privacy

The chance that the noisy released result will be R is nearly the same, whether or not you submit your information.

$$\frac{Prob(R \mid true\ world\ = D_I)}{Prob(R \mid true\ world = D_{I\pm i})} \leq e^{\varepsilon}, \qquad for\ all\ I, i, R\ and\ small\ \varepsilon > 0$$

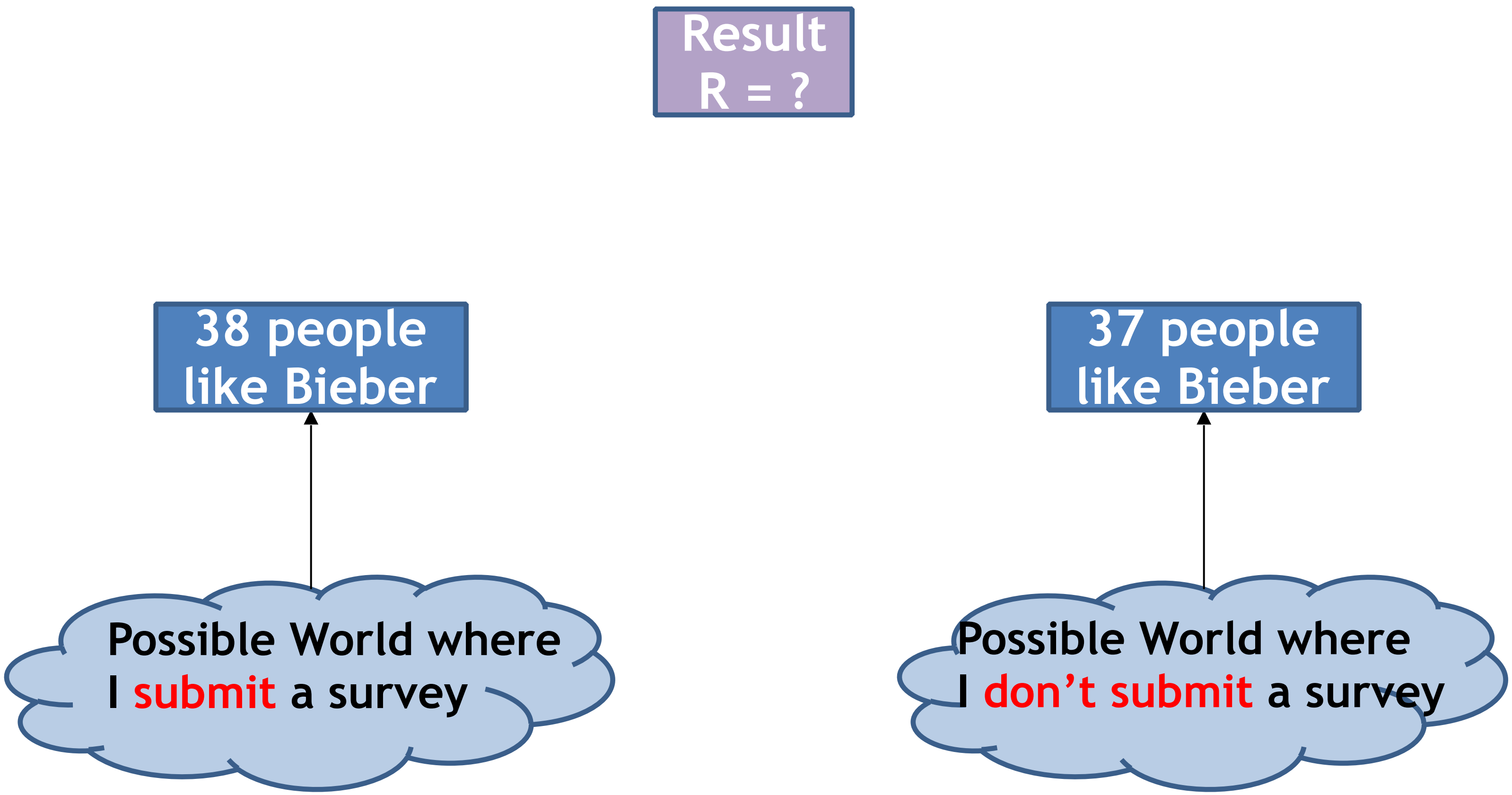
Given R, how can anyone guess which possible world it came from?



Basic Use

How do we do it?

We want to get nearly the same distribution of answers from both possible worlds. How do we bridge the gap?



Global Sensitivity

Given that **D1** and **D2** are two data sets that differ in exactly one person, and **F(D) = X** is a deterministic, non-privatized function over data set D, which returns a vector X of k real number results.

Then the ***Global Sensitivity*** of F is:

$$\Delta F = \max_{\{D1,D2\}} ||F(D1) - F(D2)||_{L1}$$

Intuitively, it's the sum of the worst case difference in answers that can be caused by adding or removing someone from a data set.

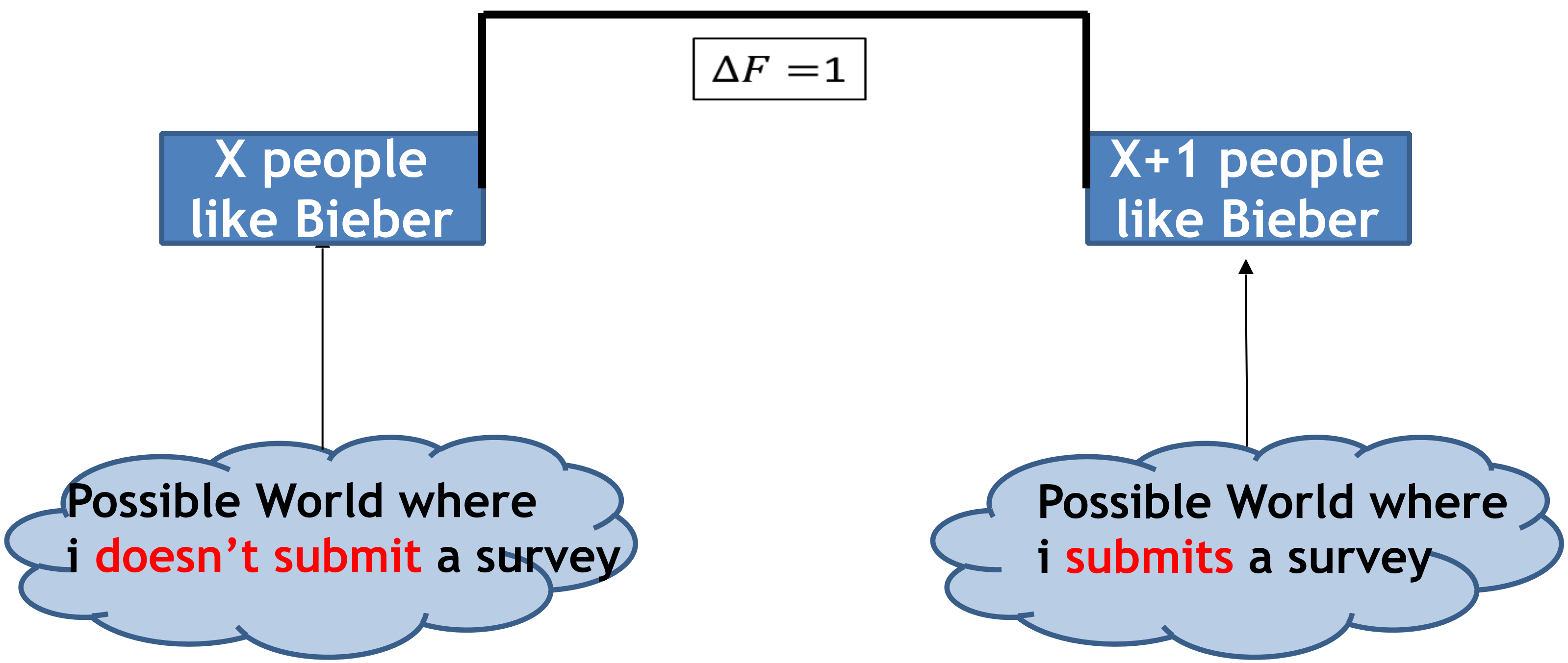
Global Sensitivity

The ***Global Sensitivity*** of F is:

$$\Delta F = \max_{\{D1,D2\}} ||F(D1) - F(D2)||_{L1}$$

Intuitively, it's the sum of the worst case difference in answers that can be caused by adding or removing someone from a data set.

How many people in the data set like Justin Bieber?



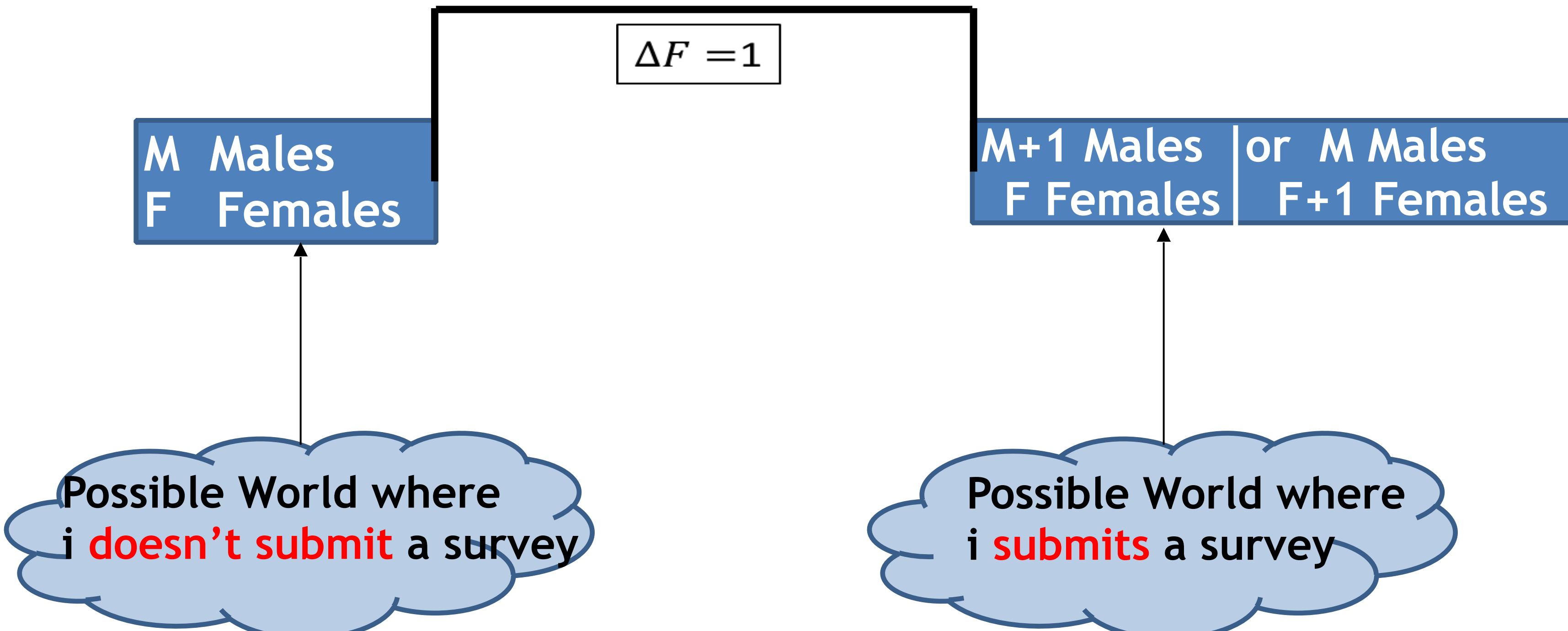
Global Sensitivity

The ***Global Sensitivity*** of F is:

$$\Delta F = \max_{\{D1,D2\}} ||F(D1) - F(D2)||_{L1}$$

Intuitively, it's the sum of the worst case difference in answers that can be caused by adding or removing someone from a data set.

How many males and females are there in the data set?



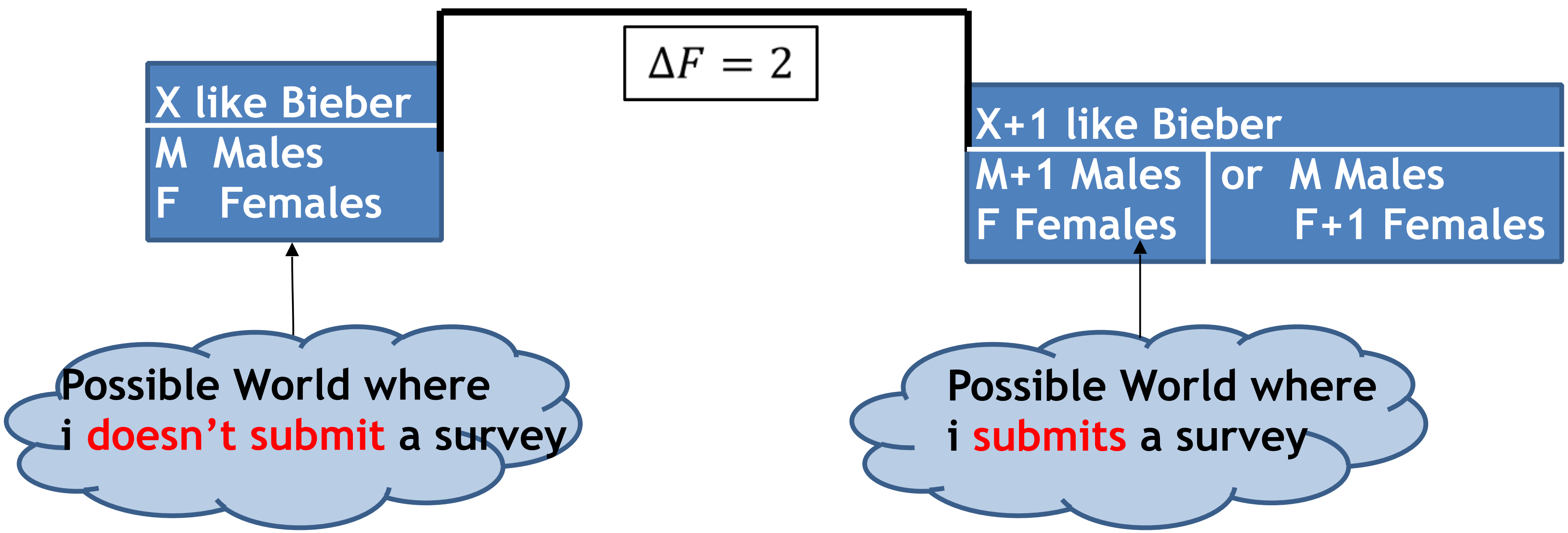
Global Sensitivity

The ***Global Sensitivity*** of F is:

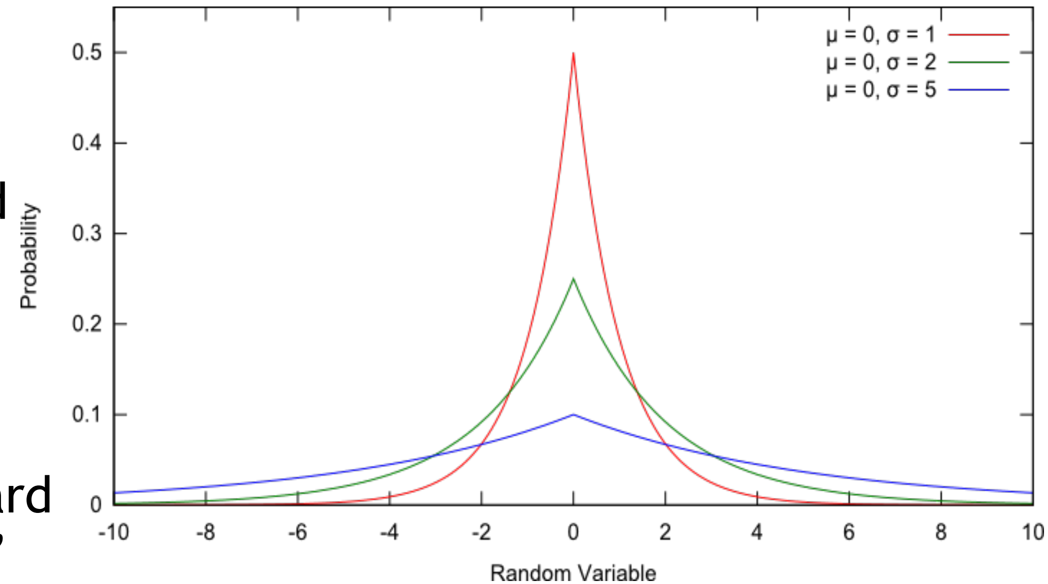
$$\Delta F = \max_{\{D1, D2\}} ||F(D1) - F(D2)||_{L1}$$

Intuitively, it's the sum of the worst case difference in answers that can be caused by adding or removing someone from a data set.

How many males and females are there in the data set?
And How many people in the data set like Justin Bieber?



Laplacian Noise



In order for our two worst-case neighboring data sets to produce a similar distribution of privatized answers, we need to add noise to span the sensitivity gap.
What noise?
Random values taken from a Laplacian distribution with standard deviation large enough to "cover" the gap. This isn't the only way to achieve differential privacy, but it's the easiest.

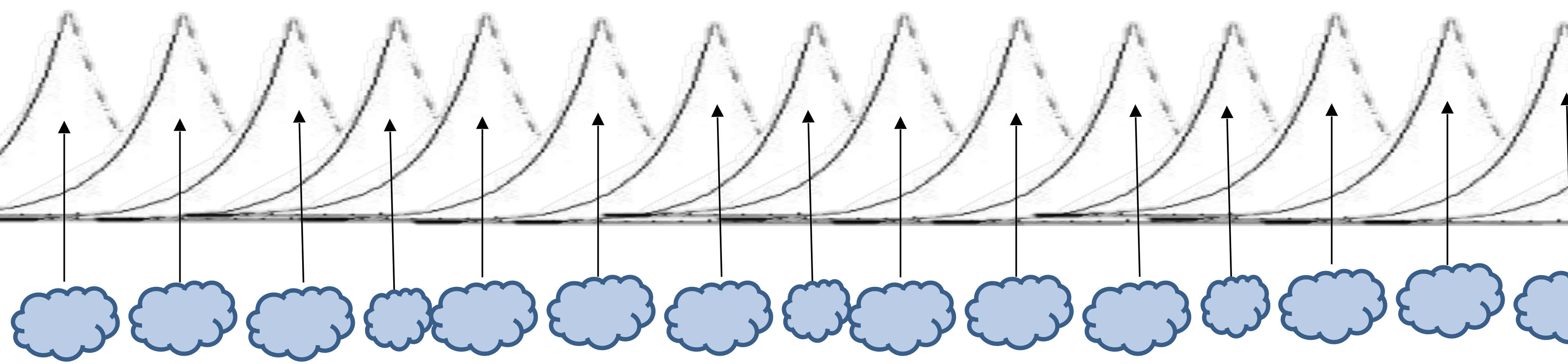
Privatizing by adding noise from the Laplacian Distribution:

$$Prob(R = x \mid D \text{ is the true world}) = \frac{\varepsilon}{2\Delta F} e^{-\frac{|x - F(D)|\varepsilon}{\Delta F}}$$

Laplacian Noise

$$Prob(R = x \mid D \text{ is the true world}) = \frac{\varepsilon}{2\Delta F} e^{-\frac{|x - F(D)|\varepsilon}{\Delta F}}$$

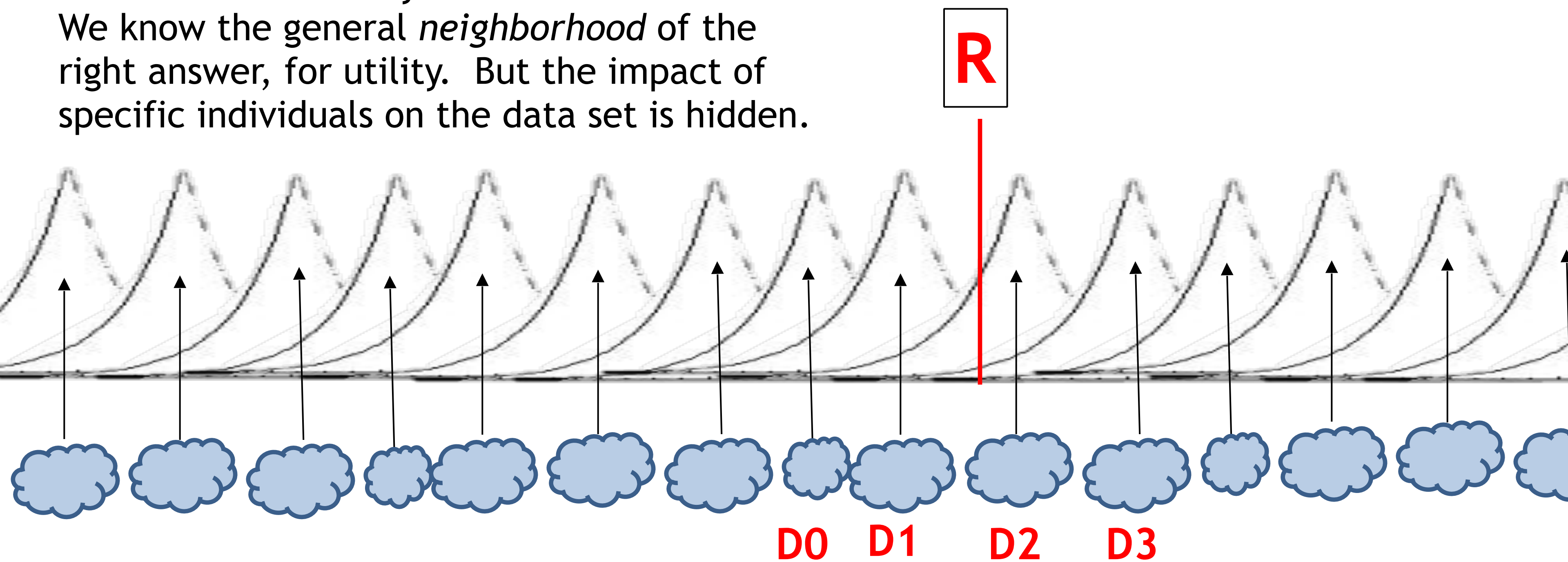
Adding Laplacian noise to the true answer means that the distribution of possible results from any data set overlaps heavily with the distribution of results from its neighbors.



Laplacian Noise

$$Prob(R = x \mid D \text{ is the true world}) = \frac{\varepsilon}{2\Delta F} e^{-\frac{|x - F(D)|\varepsilon}{\Delta F}}$$

Just by looking at the released result R, it's *very* hard to guess which world it came from and who exactly was in the data set. We know the general *neighborhood* of the right answer, for utility. But the impact of specific individuals on the data set is hidden.



Applications

Generalizing Counts

Random Forests of Binary Decision Trees: counts of randomly selected parameters are used to effectively build partitions in random decision trees.

Geetha Jagannathan, Krishnan Pillaipakkamnatt, and Rebecca N. Wright. 2009. A Practical Differentially Private Random Decision Tree Classifier. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops* (ICDMW '09). IEEE Computer Society, Washington, DC

Click Query Graphs: counts of (search query, result chosen) pairs are privatized, so search patterns can be analyzed.

Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. 2009. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web* (WWW '09). ACM, New York, NY

Beyond counting….

K-core Clustering: Individuals mapped as points in a parameter space are clustered into a reduced, robust set of points whose distribution varies little between neighboring data sets.
Dan Feldman, Amos Fiat, Haim Kaplan, and Kobbi Nissim. 2009. Private coresets. In*Proceedings of the 41st annual ACM symposium on Theory of computing* (STOC '09). ACM, New York, NY

Combinatorial Optimization: Differentially private approximation algorithms for a variety of NP-complete problems.
Anupam Gupta, Katrina Ligett, Frank McSherry, Aaron Roth, and Kunal Talwar. 2010. Differentially private combinatorial optimization. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA '10). Society for Industrial and Applied Mathematics, Philadelphia, PA

Frequent Item Set Mining: Item sets are sampled along a probability distribution which reduces the number of necessary frequency counts.
Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. 2010. Discovering frequent patterns in sensitive data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '10). ACM, New York, NY

Engineering Applications

<u>Location/Transit Data</u>: Geographical spaces are recursively partitioned using quadtrees, with areas of interest partitioned more finely.

Shen-Shyang Ho and Shuhua Ruan. 2011. Differential privacy for location pattern mining. SPRINGL '11 Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS Pages 17-24 ACM New York, NY

<u>Network Trace Analysis</u>: Counts of messages sent between network nodes are privatized and used to privately learn about network usage patterns.

Frank McSherry and Ratul Mahajan. 2010. Differentially-private network trace analysis. In*Proceedings of the ACM SIGCOMM 2010 conference* (SIGCOMM '10). ACM, New York, NY

<u>Traffic Congestion Data</u>: Streaming congestion counts at a location are sampled/estimated, privatized, post-processed to improve accuracy, and published in real time.

Fan, Liyue, and Li Xiong. "Real-time aggregate monitoring with differential privacy." In Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, 2012

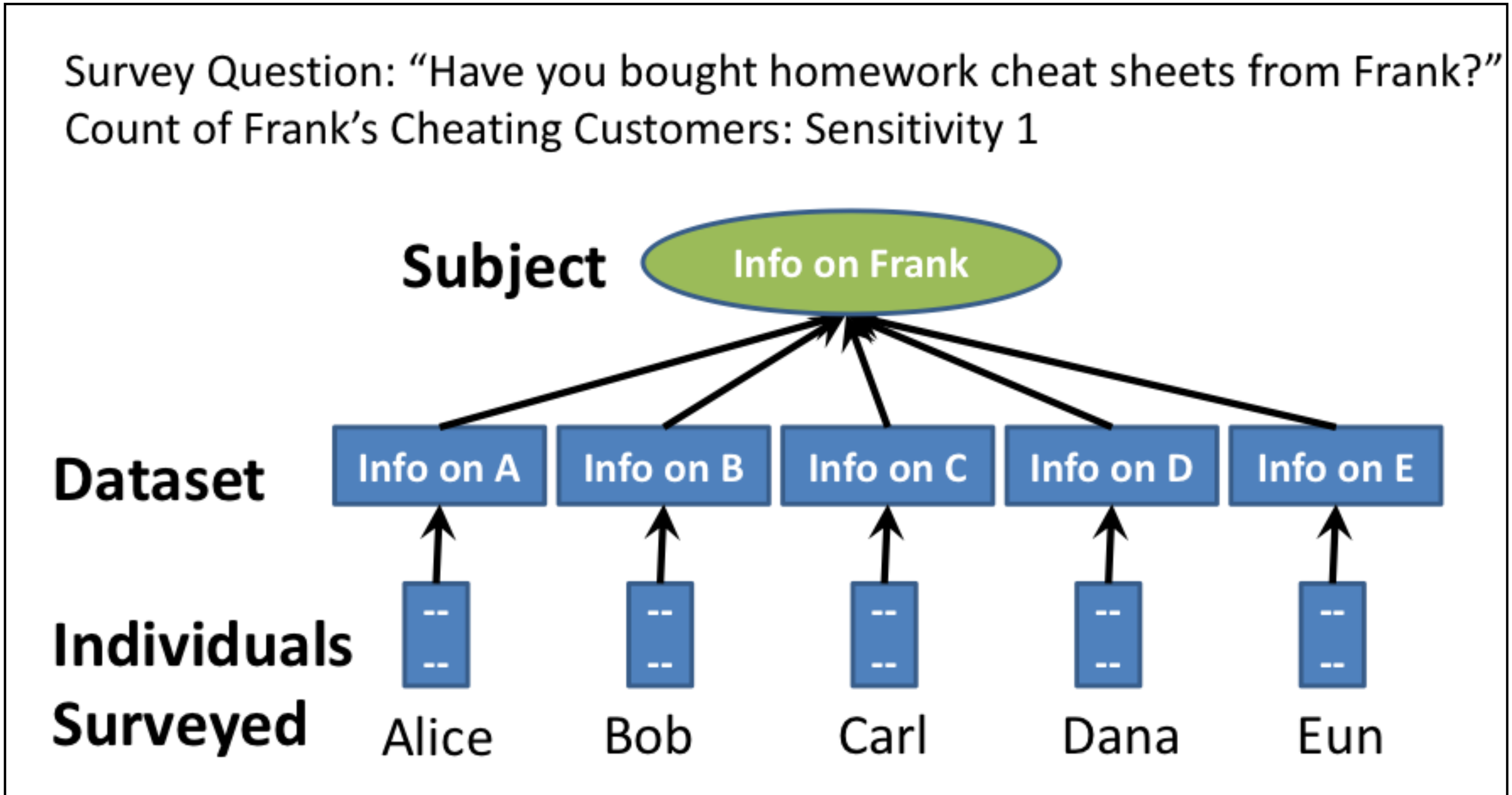
Social Network Analysis

All of the preceding work has assumed the data set was in tabular format, comprised of a list of attribute values for each individual.

Applying Differential Privacy to Social Network data, however, introduces unique challenges.

Social Network Analysis

Differential privacy protects the individuals participating in the survey, but not the subjects of the survey.

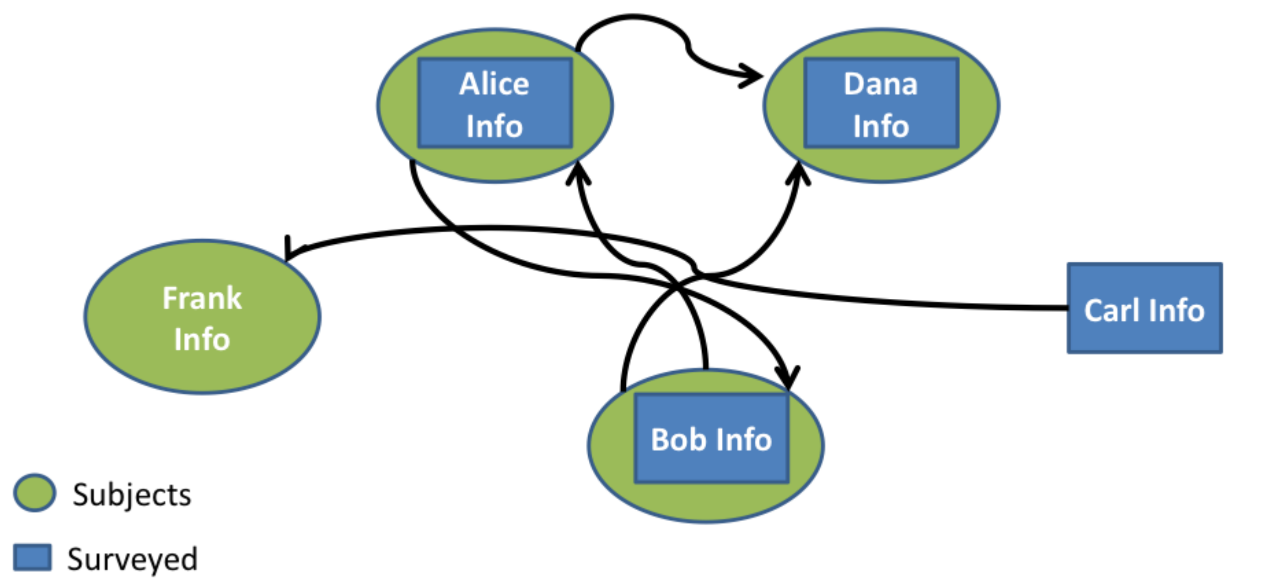


Social Network Analysis

In network data, individuals give information about each other and can be both participants and subjects of a survey.

Adapting differential privacy to networks is not straightforwards.

Name	Age	STD	Orientation	List Sexual Partners in Last 2 Years	List Friends on Campus
Alice	19	Yes	Bi	Bob, Dana, Jimmie	Dana, Eun…
Bob	20	No	Straight	Alice, Sarah…	Alice, Carl…
Carl	21	No	Gay	George, Frank	Eun, George
Dana	18	No	Straight	N/A	Alice



Social Network Analysis

Differential Privacy: Four Adaptations for Network Data

A privatized query Q satisfies ***node-privacy*** if it satisfies differential privacy for all pairs of graphs G1 = (V1, E1), G2 = (V2, E2) where V2 = V1 – x and E2 =E1 – {(v1, v2)|v1 = x V v2 = x} for x ∈ V

A privatized query Q satisfies **k-edge-privacy** if it satisfies differential privacy for all pairs of graphs G1 = (V1, E1), G2 = (V2, E2) where V1 = V2 and E2 = E1 – Ex where |Ex| = k

Social Network Analysis

Differential Privacy: Four Adaptations for Network Data

A privatized query Q satisfies ***node-privacy*** if it satisfies differential privacy for all pairs of graphs G1 = (V1, E1), G2 = (V2, E2) where V2 = V1 – x and E2 =E1 – {(v1, v2)|v1 = x ∨ v2 = x} for x ∈ V

A privatized query Q satisfies **k-edge-privacy** if it satisfies differential privacy for all pairs of graphs G1 = (V1, E1), G2 = (V2, E2) where V1 = V2 and E2 = E1 – Ex where |Ex| = k

Define PoI to be the Population of Interest, and C ⊆ PoI to be the set of people who contribute information to the data-set. A privatized query Q satisfies **contributor-privacy** if it satisfies differential privacy for all pairs of data-sets D1 = {(Info(Vi), Info(i))}, ∀i ∈ C1, and D2 = {(Info(Vi), Info(i))}, ∀i ∈ C2 where C1 = C2 - i, for some i ∈ C1.

Social Network Analysis

Differential Privacy: Four Adaptations for Network Data

A privatized query Q satisfies ***node-privacy*** if it satisfies differential privacy for all pairs of graphs G1 = (V1, E1), G2 = (V2, E2) where V2 = V1 – x and E2 =E1 – {(v1, v2)|v1 = x V v2 = x} for x ∈ V

A privatized query Q satisfies **k-edge-privacy** if it satisfies differential privacy for all pairs of graphs G1 = (V1, E1), G2 = (V2, E2) where V1 = V2 and E2 = E1 – Ex where |Ex| = k

Define PoI to be the Population of Interest, and C ⊆ PoI to be the set of people who contribute information to the data-set. A privatized query Q satisfies **contributor-privacy** if it satisfies differential privacy for all pairs of data-sets D1 = {(Info(Vi), Info(i))}, ∀i ∈ C1, and D2 = {(Info(Vi), Info(i))}, ∀i ∈ C2 where C1 = C2 - i, for some i ∈ C1.

Define a partitioned graph to be comprised of separate components such that G = {gi} for disjoint subgraphs gi. A privatized query Q satisfies **partition-privacy** if it satisfies differential privacy for all pairs of graphs G1, G2 where G1 = G2 – gi for some gi ∈ G1.

Social Network Analysis

Differential Privacy: Four Adaptations for Network Data

Contributor Privacy

Protects information contributed by one individual

Edge Privacy

Protects existence of one edge

Node Privacy

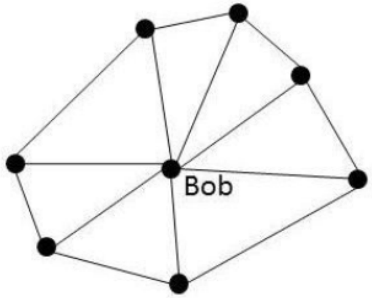
Protects existence of one node

Partition Privacy

Protects existence of one subgraph

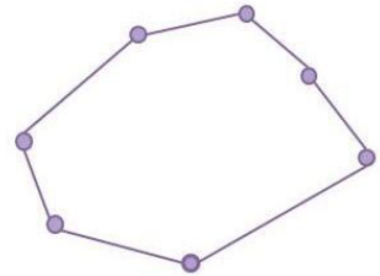
Increasing Strength of Privacy Guarantee

Social Network Analysis



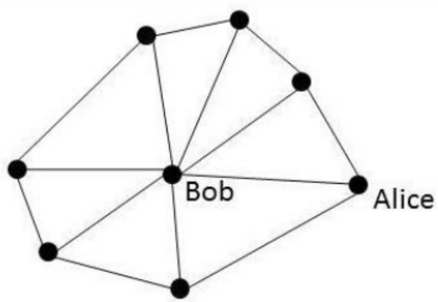
Distribution with Bob

Count	0	7	...	1
Degree	2	3	...	7



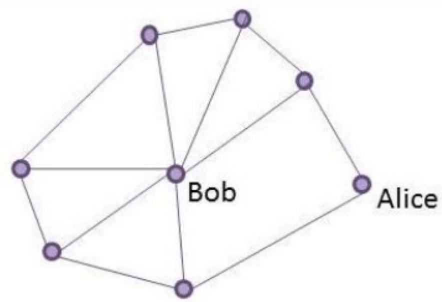
Distribution without Bob

Count	7	0	...	0
Degree	2	3	...	7



True Distribution

Count	0	7	...	0	1
Degree	2	3	...	6	7



Without Bob-Alice Friendship

Count	1	6	...	1	0
Degree	2	3	...	6	7

Degree Distribution

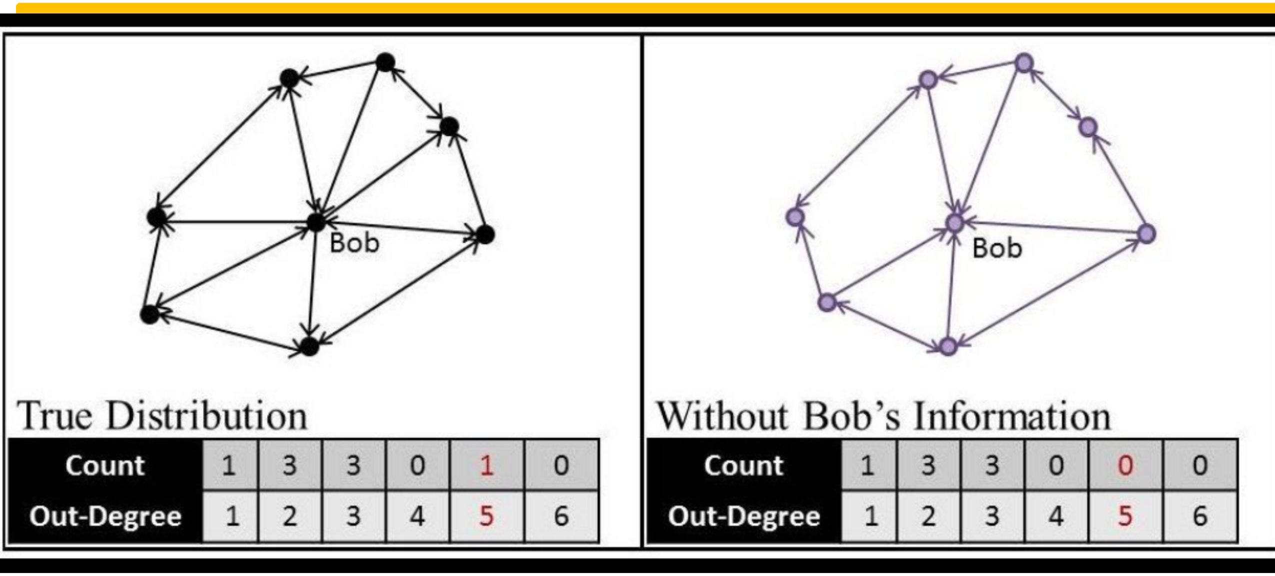
Node Privacy:
ΔF is unbounded

Global sensitivity is unbounded if d is unbounded.

Edge Privacy:
$\Delta F = 2$

Post-processing may improve results.
[Hay 2009]

Social Network Analysis



True Distribution

Count	1	3	3	0	1	0
Out-Degree	1	2	3	4	5	6

Without Bob's Information

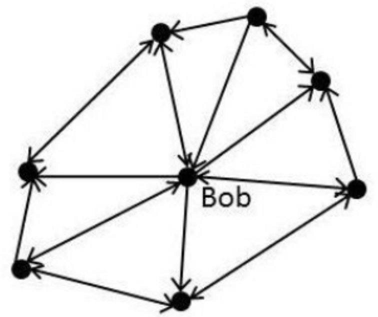
Count	1	3	3	0	0	0
Out-Degree	1	2	3	4	5	6

Degree Distribution

Contributor Privacy:
$\Delta F = 1$

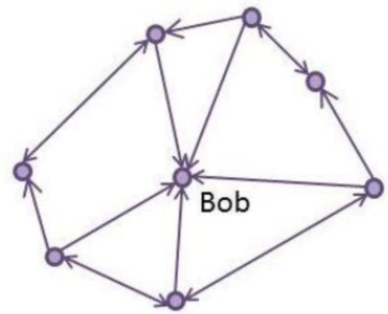
Degree distribution records nodes' *perceived* friend count (out-degree).

Social Network Analysis



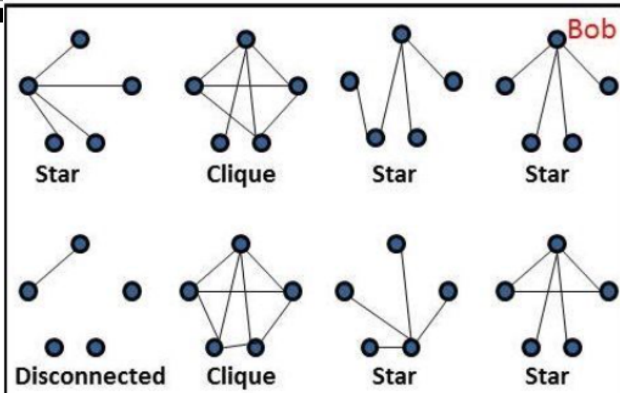
True Distribution

Count	1	3	3	0	1	0
Out-Degree	1	2	3	4	5	6



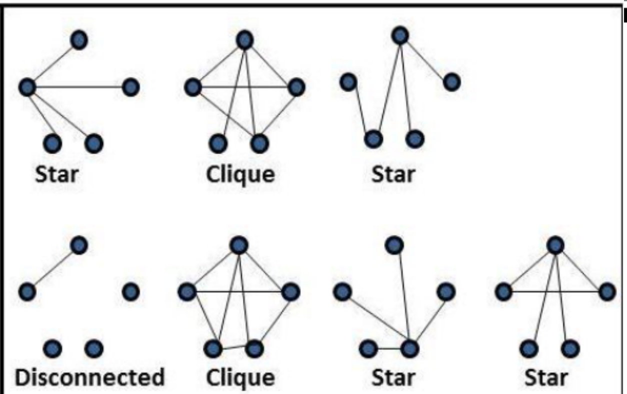
Without Bob's Information

Count	1	3	3	0	0	0
Out-Degree	1	2	3	4	5	6



Cliques	Stars	Disconnected
2	5	1

True Distribution of Working-groups



Cliques	Stars	Disconnected
2	4	1

Distribution Without Bob's Group

Degree Distribution

Contributor Privacy:
$\Delta F = 1$

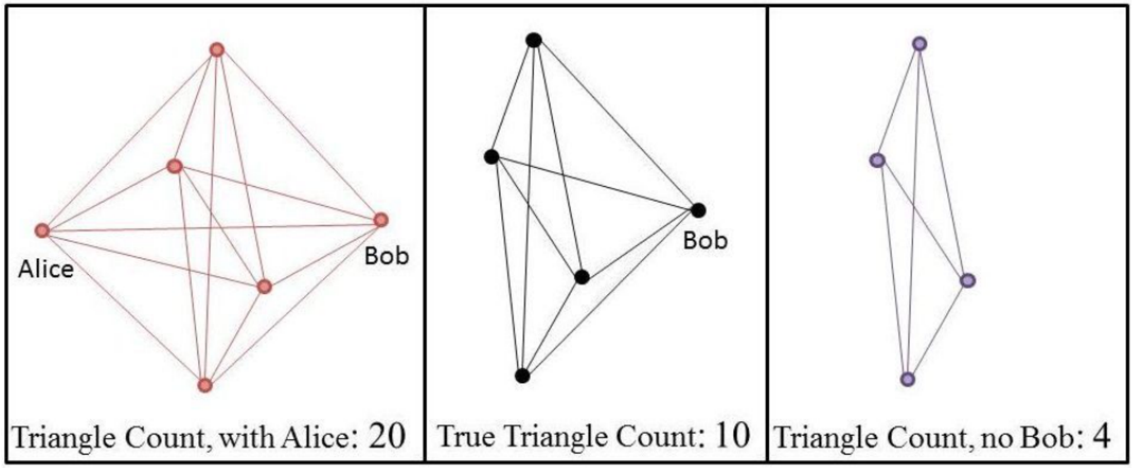
Degree distribution records nodes' *perceived* friend count (out-degree).

Partition Privacy:
$\Delta F = 1$

Histogram records degree distribution types over collection of disjoint graphs.

Social Network Analysis

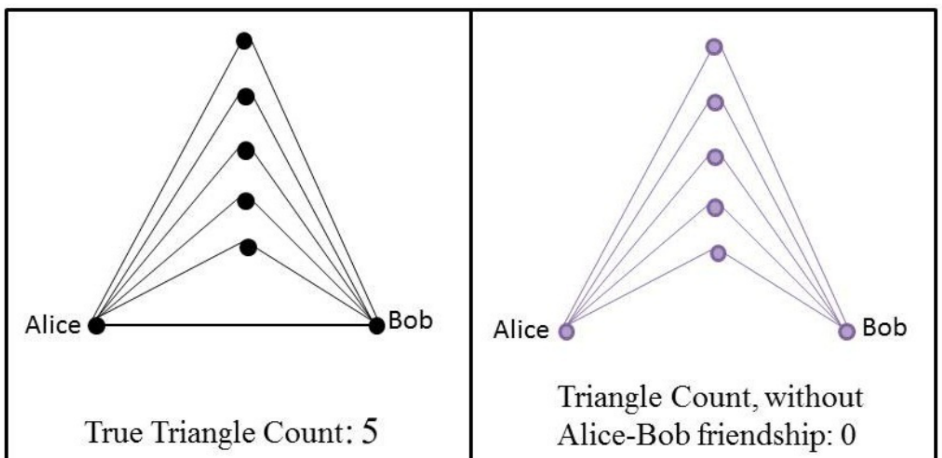
Triangle Counts



Triangle Count, with Alice: 20 | True Triangle Count: 10 | Triangle Count, no Bob: 4

<u>Node Privacy</u>:
ΔF is unbounded

Global sensitivity is unbounded if d is unbounded.
Smooth sensitivity is bounded, but quite high: O(d^2) [Blocki 2012]



True Triangle Count: 5 | Triangle Count, without Alice-Bob friendship: 0

<u>Edge Privacy</u>:
ΔF is unbounded

Global sensitivity is unbounded if d is unbounded.
Smooth sensitivity is bounded, but added noise can be very high: 10*T [Karwa 2011]

Social Network Analysis

Differential Privacy: Triangle-Count, Clustering Coefficient

Algorithm 1 A survey gathering information about triangles.

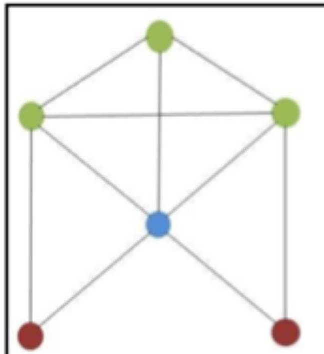
function TRIANGLEQUERY
- $friendlist \leftarrow$ Query("Who are your friends?")
- $friendpairs \leftarrow$ CrossProduct($friendlist, friendlist$)
- $outdegree \leftarrow$ Size($friendlist$)
- $triangles \leftarrow$ Query("Which of these pairs are friends with each other?", $friendpairs$)
- $trianglecount \leftarrow$ Size($triangles$)
- **return** $(outdegree, trianglecount)$

end function

Algorithm 2 Privatizing local clustering coefficient distribution data.

function PRIVATECLUSTERING($deg_{low}, deg_{med}, data$)
- Initialize($bins[][]$)
- **for all** $(nodeDegree, triangleCount) \in data$ **do**
  - $degBin \leftarrow$ Partition($nodeDegree, deg_{low}, deg_{med}$)
  - $localCluster \leftarrow triangleCount/(nodeDegree * (nodeDegree - 1))$
  - $triBin \leftarrow$ Partition($localCluster, 1/3, 2/3$)
  - $bin[degBin][triBin] \leftarrow bin[degBin][triBin] + 1$
- **end for**
- **for** $i = 0 \rightarrow 2, j = 0 \rightarrow 2$ **do**
  - $bins[i][j] \leftarrow bins[i][j] +$ LaplacianNoise(1)
- **end for**
- **return** $bins$

end function



Clustering Coefficient Distribution

Clustering Coefficient	Node Degree: Low	Node Degree: Med	Node Degree: High
Low	2	0	0
Med	0	3	0
High	0	1	0

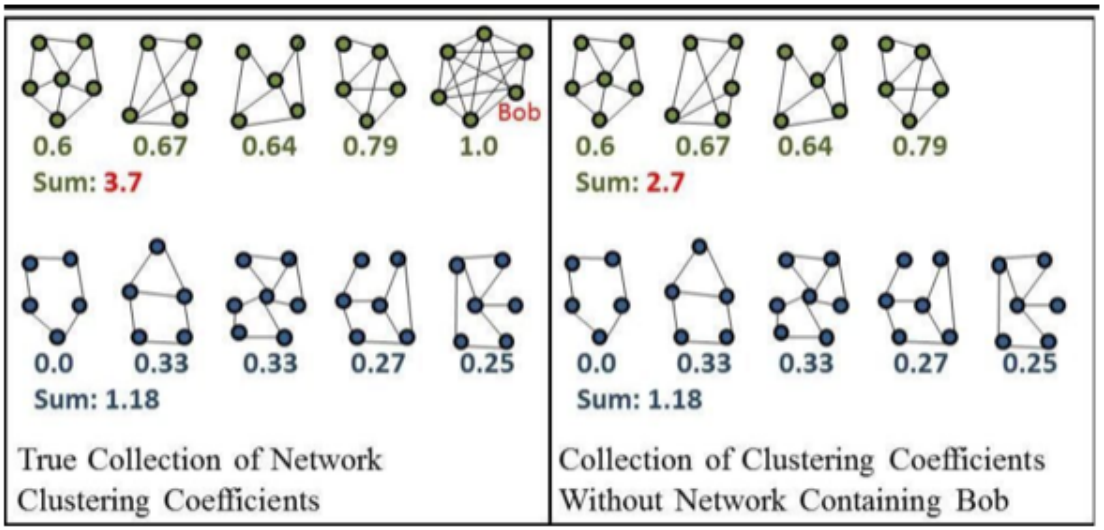
Contributor Privacy:

$\Delta F = 1$

Two-dimensional histogram records nodes' *perceived* clustering coefficient and degree.

Social Network Analysis

Differential Privacy: Triangle-Count, Clustering Coefficient



Partition Privacy:
$\Delta F = 1$

Compute and compare average global clustering coefficients across sets of graphs.

An example comparison of graphs sets by average global clustering coefficients:

MaleDormsClustering = M/N1 where M = $\Sigma_{G \in \text{MaleDorms}}$ *clustering-coefficient*(G), N1 = |MaleDorms|

FemaleDormsClustering = F/N2 where F = $\Sigma_{G' \in \text{FemaleDorms}}$ *clustering-coefficient*(G'), N2 = |FemaleDorms|

If N1, N2 are publicly known then the sensitivity of these means is equal to the sensitivity of the numerators. Since *range(clustering-coefficient())* = [0,1], the sensitivity of the numerator is 1.
The expected noise value added to the function result is: Lap(1/epsilon)/N

Learning Analytics

Big Data in Education

- Gained attention with school accountability testing, MOOCS, and ubiquitous smart phones.
- **Objectives**: Predict student success, improve instruction, improve assessment, improve convenience of data management
- **Data Sources**: Everything. Grades, tests, surveys, homework, attendance, forum posts, chat logs, data collected from interaction with apps. Social networks, disposition/mood, content analysis, attention, even neural data.
- **Academic Research**: Joins Machine Learning, HCI, NLP, Education, Psychology, and others.
- **Tech**: Enormous and growing market of software, apps, cloud services.
- **Policy**: Tech-funded lobbying groups like the Data Quality Campaign set state goals like career-long ID#'s for students.

Learning Analytics

Legalities:

- **FERPA (before 2008):** Data access limited to teachers and school officials.
- **FERPA (after 2008):** Access increased to include: "contractors, consultants, volunteers, and other outside parties providing institutional services and functions".
- **Breach Disclosure:** Because educational data companies are not storing financial data, they may not be legally required to disclose leaks that occur.
- **FERPA trumps HIPAA:** Student health records submitted to a school are no longer covered by HIPAA

Learning Analytics

A Compromise

- Leave raw data on school district owned computers, and provide (mandate) good security.
- Share ***aggregated*** (not simply anonymized), ***privatized*** data for analysis.
- Feel unafraid of leaks. Leaked data might lose its financial value for the company that held it, but it cannot with high probability be used to victimize individual students.
- **How?:** We'll demonstrate how this could be done for a representative set of example use-cases inspired by learning analytics literature and real-life tools.

Learning Analytics

Example Use Case 1: Sentiment Analysis

A school district's administration uses a cloud-based sentiment analysis tool to get feedback on the impact of various curriculum choices.

Data including student reading assignment responses and homework help forums is analyzed in order to understand students' reaction to various topics, such as assigned books or math concepts.

In each segment, the cloud-based service will identify words indicating positive and negative affect, topic words, and authorship. This data will then be used to create a report for the district administration, including a compact visualization of their students' response to the curriculum.

This report is valuable as it provides immediate, authentic feed-back which may be difficult to achieve through traditional surveys or student evaluation forms.

Learning Analytics

Sentiment Analysis:

Math Help Forum
Carla: Did anyone get problem six on the homework? I'm **lost**. I **hate exponents**.
Alice: I think **I got it**. My mom showed me this trick that makes it **easy**.
Bob: Really? Cool! I'm **totally lost** on that section too. Do you want to meet up to study?
Carla: Yah
Alice: Ok sure. Meet at my house after school? It's **214 Elm St**, just behind the krogers. U can call me if u need. **614 123-4567**
Bob: Ok. Thanks **Alice**!

Alice Howard
Chapter 2 Response:
This chapter talked about how Elise ran away from home after her dad came home drunk and got into a fight with her mom. I **liked** this chapter because I think the author did a **very good** job of describing the characters and scenes. I also **really liked** how Elise was brave enough to run away. **Before my mom divorced my dad, they would fight like that sometimes, but I never ran away**. I am **looking forwards** to reading what happens next.

Example Raw Data Fields

- FULL NAME
- STUDENT ID#
- STUDENT CONTACT INFO
- RESPONSE ESSAYS/FORUM POSTS

Potential Privacy Invasions

Student Address, ID# (possibly SSN)

Any Private Information Revealed In Text.

Learning Analytics

Example Of Aggregated Data (Hypothetical)

Topic vs. Affect (Count of individuals showing positive or negative affect on a given week's forum posts.)

Topic	3/12-3/18	Positive Affect	Negative Affect
	Exponents	3	16
	Percentages	6	12
	Measurement	14	4
	Inequalities	10	2

Application of Differential Privacy: If a student deletes one week of their text segments covering one topic, or writes new text segments covering a new topic, then at most one of the aggregate counts will be affected by at most |1|. Thus the global sensitivity of a student's affect regarding a topic (each week) is 1, and we can provide differential privacy protection by adding laplacian noise to the aggregated data set with $\Delta F = 1$, epsilon = $\sqrt{2}$.

Learning Analytics

Example Of Aggregated Data (Hypothetical)

Topic vs. Affect (Count of individuals showing positive or negative affect on a given week's forum posts.)

Topic: 3/12-3/18	Positive Affect	Negative Affect
Exponents	3	16
Percentages	6	12
Measurement	14	4
Inequalities	10	2

Example Of Privatized Data (Hypothetical)

Topic vs. Affect (Count of individuals showing positive or negative affect on a given week's forum posts.)

Topic: 3/12-3/18	Positive Affect	Negative Affect
Exponents	2.4	18.2
Percentages	4.1	10.2
Measurement	15.2	3.8
Inequalities	11.7	2.4

Questions?