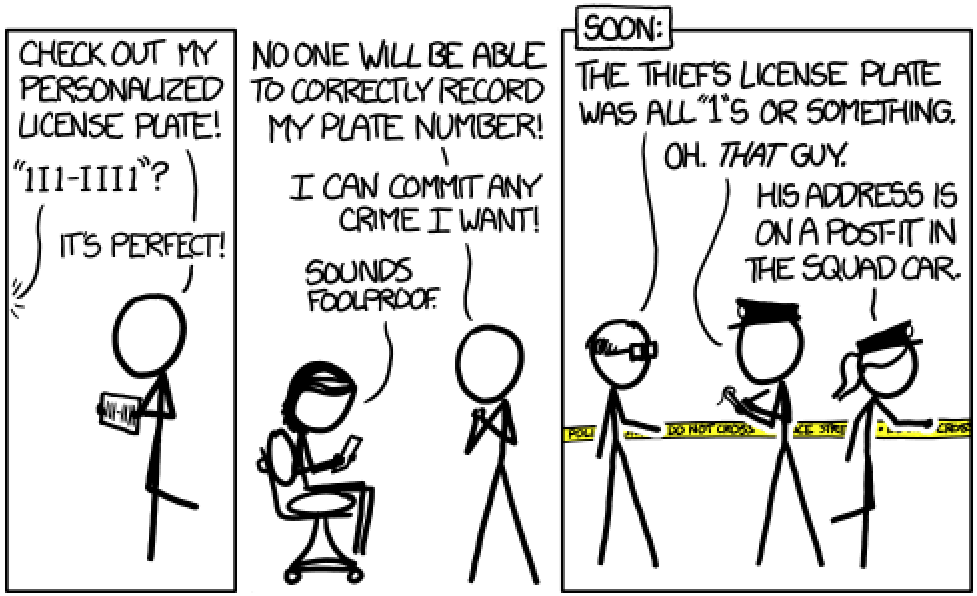


Privacy Definitions: Beyond Anonymity

CompSci 590.03
Instructor: Ashwin Machanavajjhala

Duke UNIVERSITY

Announcements

- Some new project ideas added

- Please meet with me at least once before you finalize your project (deadline Sep 28).

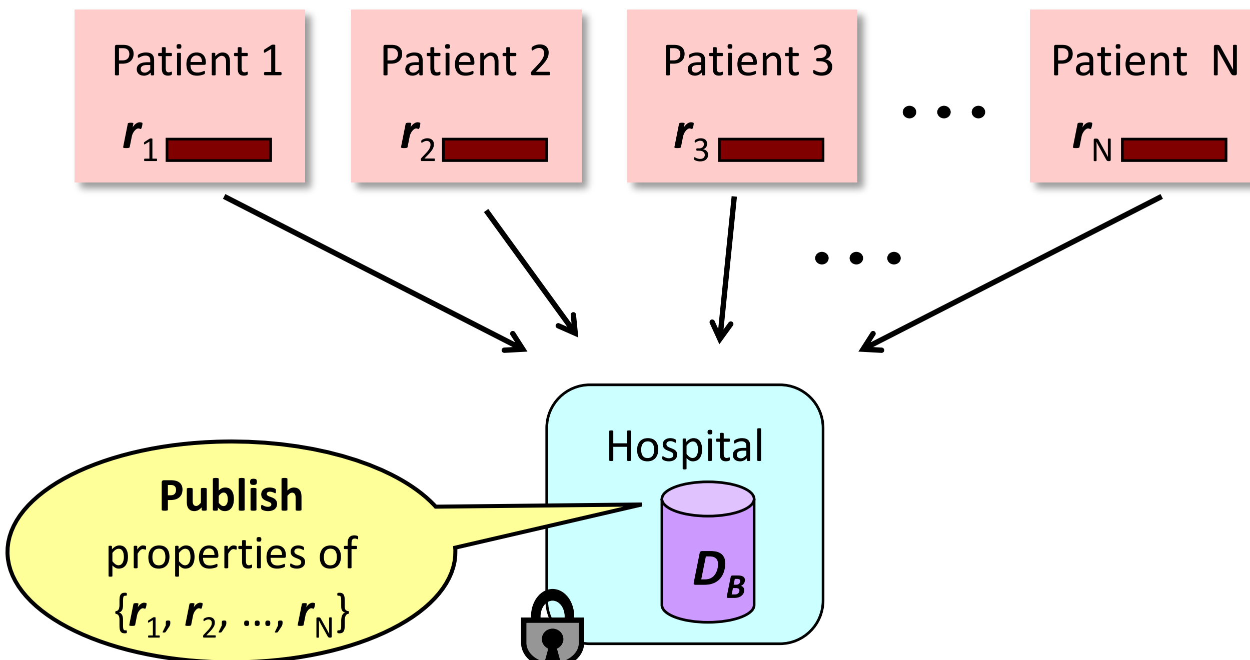
Outline

- Does k-anonymity guarantee privacy?
- L-diversity
- T-closeness

Data Publishing

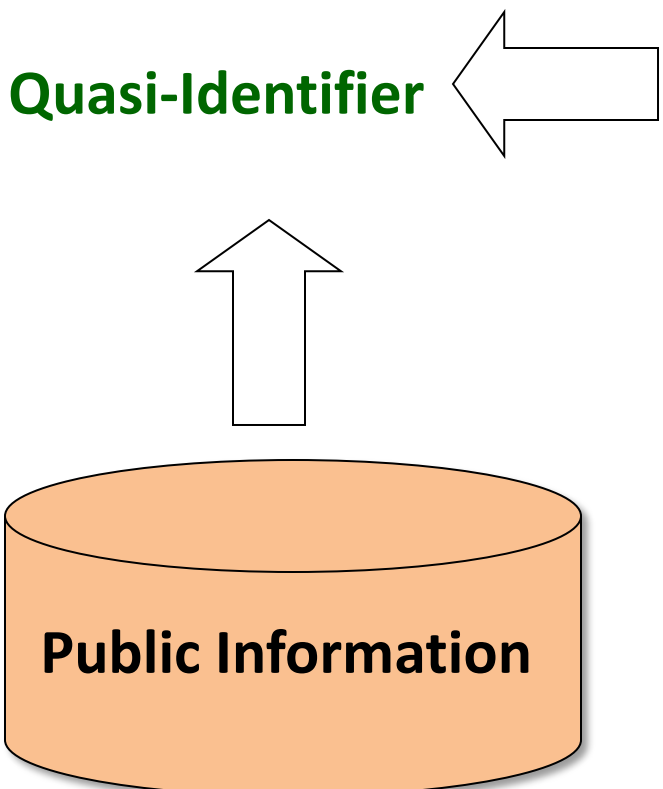
Publish information that:

- Discloses as much statistical information as possible.
- Preserves the privacy of the individuals contributing the data.



Duke UNIVERSITY

Privacy Breach: linking identity to sensitive info.



Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

Duke UNIVERSITY

k-Anonymity using Generalization

Quasi-identifiers (Q-ID)

can identify individuals in the population

table T* is k-anonymous

if each

```
SELECT COUNT(*)
FROM T*
GROUP BY Q-ID
```

is ≥ k

Parameter k indicates “degree” of anonymity

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

k-Anonymity: A popular privacy definition

Complexity

- k-Anonymity is NP-hard
- *(log k)* Approximation Algorithm exists

Algorithms

- Incognito (use monotonicity to prune generalization lattice)
- Mondrian (multidimensional partitioning)
- Hilbert (convert multidimensional problem into a 1d problem)
- …

Does k-Anonymity guarantee sufficient privacy ?

Attack 1: Homogeneity

Bob has Cancer

Name	Zip	Age	Nat.
Bob	13053	35	??

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Duke UNIVERSITY

Attack 2: Background knowledge

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Attack 2: Background knowledge

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Japanese have a very low incidence of Heart disease.

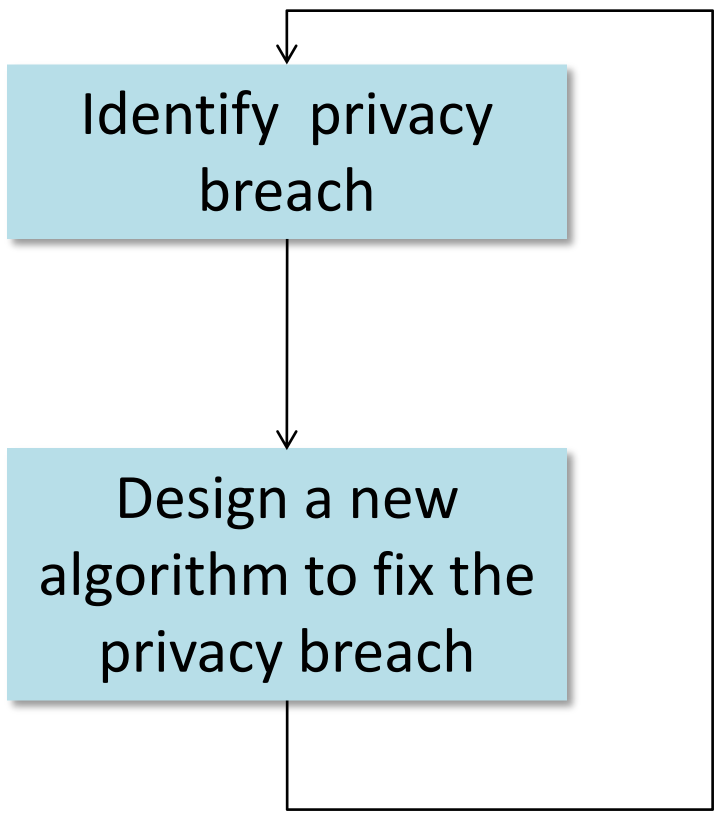
Umeko has Flu

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Duke UNIVERSITY

Q: How do we ensure the privacy of published data?

Method 1:
Breach and Patch



The MA Governor Breach and the AOL Privacy Breach caused by **re-identifying individuals.**

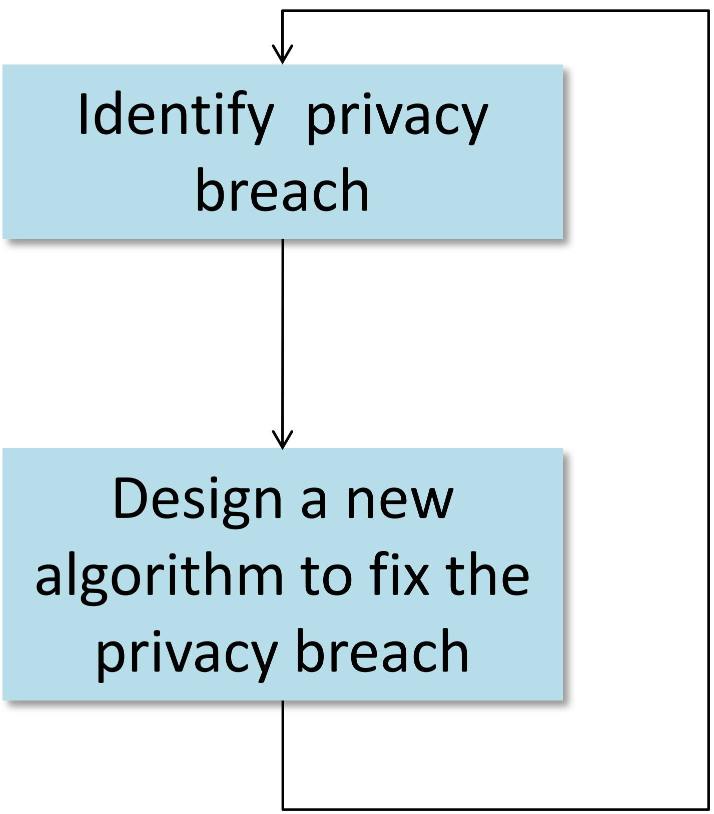
k-Anonymity only considers the risk of re-identification.

Adversaries with background knowledge can breach privacy even without re-identifying individuals.

Duke UNIVERSITY

Limitations of the Breach and Patch methodology.

Method 1:
Breach and Patch



1. **A data publisher may not be able to enumerate all the possible privacy breaches.**
2. **A data publisher does not know what other privacy breaches are possible.**

Q: How do we ensure the privacy of published data?

Method 1:
Breach and Patch

Identify privacy breach

Design a new algorithm to fix the privacy breach

Method 2:
Define and Design

Formally specify the privacy model

Derive conditions for privacy

Design an algorithm that satisfies the privacy conditions

Recall the attacks on k-Anonymity

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Japanese have a very low incidence of Heart disease.

Umeko has Flu

Name	Zip	Age	Nat.
Bob	13053	35	??

Bob has Cancer

Zip	Age	Nat.	Disease
130**	<30	*	~~Heart~~
130**	<30	*	~~Heart~~
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Duke UNIVERSITY

3-Diverse Table

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Japanese have a very low
in

L-Diversity Principle:
Every group of tuples with the same Q-ID values has ≥ L distinct sensitive values of roughly equal proportions.

Zip	Age	Nat.	Disease
1306*	<=40	*	~~Heart~~
1306*	<=40	*	Flu
1306*	<=40	*	Cancer
	<=40	*	Cancer
	>40	*	Cancer
	>40	*	Heart
	>40	*	Flu
	>40	*	Flu
1305*	<=40	*	Heart
1305*	<=40	*	Flu
1305*	<=40	*	Cancer
1305*	<=40	*	Cancer

Name	Zip	Age	Nat.
Bob	13053	35	??

Bob has ??

Duke UNIVERSITY

L-Diversity: Privacy Beyond K-Anonymity

[Machanavajjhala et al ICDE 2006]

L-Diversity Principle:

Every group of tuples with the same Q-ID values has ≥ L distinct "well represented" sensitive values.

Questions:

- What kind of adversarial attacks do we guard against?
- Why is this the right definition for privacy?
  - What does the parameter L signify?

Duke UNIVERSITY

Method 2:
Define and Design

Formally specify the privacy model

↓

Derive conditions for privacy

↓

Design an algorithm that satisfies the privacy conditions

1. Which information is sensitive?
2. What does the adversary know?
3. How is the disclosure quantified?

- **L-Diversity**

- **L-Diverse Generalization**

Duke UNIVERSITY

Privacy Specification for L-Diversity

- The link between identity and attribute value is the sensitive information.

  "Does Bob have Cancer? Heart disease? Flu?"

  "Does Umeko have Cancer? Heart disease? Flu?"

- Adversary knows ≤ *L-2* negation statements.

  "Umeko does not have Heart Disease."

  – Data Publisher may not know exact adversarial knowledge

- Privacy is breached … ribute value with high probabilit…

  Pr[*"Bob has Cancer"* | *published table, adv. knowledge*] > t

Individual *u* does not have a specific disease *s*

Method 2:
Define and Design

Formally specify the privacy model

↓

Derive conditions for privacy

↓

Design an algorithm that satisfies the privacy conditions

1. Which information is sensitive?
2. What does the adversary know?
3. How is the disclosure quantified?

- **L-Diversity**

- **L-Diverse Generalization**

Calculating Probabilities

Set of all possible worlds

Every world represents a unique assignment of diseases to individuals

	World 1	World 2	World 3	World 4	World 5
Sasha	Cancer	Heart	Heart	Flu	Heart
Tom	Cancer	[illegible]	Flu	Heart	Flu
Umeko	Cancer	[illegible]	Flu	Heart	Heart
Van	[illegible]	[illegible]	Heart	Flu	Flu
Amar	[illegible]	[illegible]	Heart	Cancer	Flu
Boris	[illegible]	[illegible]	Cancer	Flu	Heart
Carol	[illegible]	[illegible]	Flu	Heart	Flu
Dave	Cancer	Flu	Flu	Flu	Cancer
Bob	Cancer	Cancer	Cancer	Cancer	Cancer
Charan	Cancer	Cancer	Cancer	Cancer	Cancer
Daiki	Cancer	Cancer	Cancer	Cancer	Cancer
Ellen	Cancer	Cancer	Cancer	Cancer	Cancer

…

Calculating Probabilities

Set of all possible worlds with T*

	T*	World 1	World 2	World 3	World 4	World 5
Sasha	**Cancer 0**	Cancer	Heart	Heart	Flu	Heart
Tom	**Heart 2**	Cancer	Heart	Flu	Heart	Flu
Umeko	**Flu 2**	Cancer	Flu	Flu	Heart	Heart
Van		Cancer	Flu	Heart	Flu	Flu
Amar	**Cancer 1**	Cancer	Cancer	Heart	Cancer	Flu
Boris	**Heart 1**	Cancer	Heart	Cancer	Flu	Heart
Carol	**Flu 2**	Cancer	Flu	Flu	Heart	Flu
Dave		Cancer	Flu	Flu	Flu	Cancer
Bob		Heart	Cancer	Cancer	Cancer	Cancer
Charan	**Cancer 4**	Flu	Cancer	Cancer	Cancer	Cancer
Daiki	**Heart 0**	Cancer	Cancer	Cancer	Cancer	Cancer
Ellen	**Flu 0**	Cancer	Cancer	Cancer	Cancer	Cancer

…

Duke UNIVERSITY

Pr[Umeko has Flu| B, T*] =

worlds consistent ... has Flu

worlds ... ent with T*

Set of worlds consistent with B, T*

	T*	
Sasha	Cancer	0
Tom	Heart	2
Umeko	Flu	2
Van		
Amar	Cancer	1
Boris	Heart	1
Carol	Flu	2
Dave		
Bob	Cancer	4
Charan	Heart	0
Daiki	Flu	0
Ellen		

World 2	World 3	World 4	World 5
Heart	Heart	Flu	Heart
Heart	Flu	Heart	Flu
Flu	Flu	Heart	Heart
Flu	Heart	Flu	Flu
Cancer	Heart	Cancer	Flu
Heart	Cancer	Flu	Heart
Flu	Flu	Heart	Flu
Flu	Flu	Flu	Cancer
Cancer	Cancer	Cancer	Cancer
Cancer	Cancer	Cancer	Cancer
Cancer	Cancer	Cancer	Cancer
Cancer	Cancer	Cancer	Cancer

B: Umeko.Disease ≠ Heart

...

Pr[Umeko has Flu | B, T*] =

$$\frac{\text{\# worlds consistent with B, T* where Umeko has Flu}}{\text{\# worlds consistent with B, T*}}$$

T*

Sasha Tom **Umeko** Van	**Cancer**	**0**
	Heart	**2**
	Flu	**2**
Amar Boris Carol Dave	**Cancer**	**1**
	Heart	**1**
	Flu	**2**
Bob Charan Daiki Ellen	**Cancer**	**4**
	Heart	**0**
	Flu	**0**

Counting the # worlds consistent with B, T* is tedious.

(and is intractable for more complex forms of B)

B: Umeko.Disease ≠ Heart

Duke UNIVERSITY

Pr[Umeko has Flu | B, T*] =

$$\frac{\text{\# worlds consistent with B, T* where Umeko has Flu}}{\text{\# worlds consistent with B, T*}}$$

T*

Sasha Tom **Umeko** Van	**Cancer**	**0**
	Heart	**2**
	Flu	**2**
Amar Boris Carol Dave	**Cancer**	**1**
	Heart	**1**
	Flu	**2**
Bob Charan Daiki Ellen	**Cancer**	**4**
	Heart	**0**
	Flu	**0**

Theorem:

worlds consistent with B, T* where Umeko has Flu is

proportional to

tuples in Umeko's group who have Flu.

B: Umeko.Disease ≠ Heart

Data publisher does not know the adversary's knowledge about *u*

- *Different adversaries have varying amounts of knowledge.*
- *Adversary may have different knowledge about different individuals.*

Therefore, in order for privacy,
check for each individual *u*, and each disease *s*

Pr[*"u has disease s"* | *T*,* *adv. knowledge about u*] < t

And we are done … ??

NO

L-Diversity:
Guarding against unknown adversarial knowledge.

- Limit adversarial knowledge
  - Knows ≤ *(L-2)* negation statements of the form *"Umeko does not have a Heart disease."*
- Consider the worst case
  - Consider all possible conjunctions of ≤ *(L-2)* statements

At least L sensitive values should appear in every group

L = 5

Cancer	**10**
~~**Heart**~~	~~**5**~~
~~**Hepatitis**~~	~~**2**~~
~~**Jaundice**~~	~~**1**~~

Pr[Bob has Cancer] = 1

Duke UNIVERSITY

Guarding against unknown adversarial knowledge

- Limit adversarial knowledge
  - Knows ≤ *(L-2)* negation statements of the form *"Umeko does not have a Heart disease."*
- Consider the worst case
  - Consider all possible conjunctions of ≤ *(L-2)* statements

The L distinct sensitive values in each group should be roughly of equal proportions

L = 5

Cancer	**1000**
~~**Heart**~~	~~**5**~~
~~**Hepatitis**~~	~~**2**~~
~~**Jaundice**~~	~~**1**~~
Malaria	**1**

Pr[Bob has Cancer] ≈ 1

Guarding against unknown adversarial knowledge

L = 5

The L distinct sensitive values in each group should be roughly of equal proportions

Cancer	**1000**
~~**Heart**~~	~~**5**~~
~~**Hepatitis**~~	~~**2**~~
~~**Jaundice**~~	~~**1**~~
Malaria	**1**

Pr[Bob has Cancer] ≈ 1

Let *t = 0.75.* Privacy of individuals in the above group is ensured if ,

$$\frac{\#\ \text{Cancer}}{\#\ \text{Cancer} + \#\ \text{Malaria}} < 0.75$$

Duke UNIVERSITY

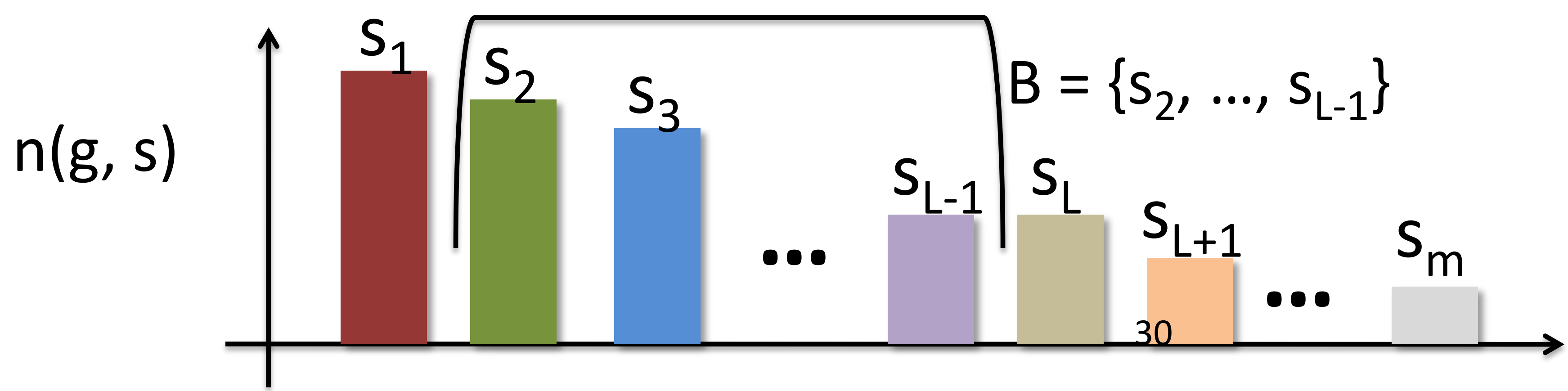
Theorem:

For all groups *g*, for all *s in S*, and for all B, |B| ≤ (L-2)

$$\frac{n(g, s)}{\sum_{s' \varepsilon (S \backslash B)} n(g, s')} \leq t$$

is equivalent to

$$\frac{n(g, s_1)}{n(g, s_1) + n(g, s_L) + n(g, s_{L+1}) + \ldots + n(g, s_m)} \leq t$$



Method 2:
Define and Design

Formally define privacy

↓

Derive conditions for privacy

↓

Design an algorithm that matches privacy conditions

1. Which information is sensitive?
2. What does the adversary know?
3. How is the disclosure quantified?

- **L-Diversity**

- **L-Diverse Generalization**

Algorithms for L-Diversity

- Checking whether T* is L-Diverse is straightforward
  - In every group g,
  - Check the L-Diversity condition.

- Finding an L-Diverse table is a Lattice search problem (NP-complete)

Algorithms for L-Diversity

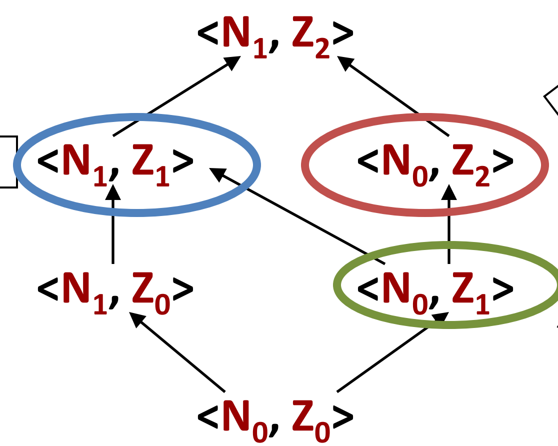
- Finding an L-Diverse table is a Lattice search problem (NP-complete)

Q =

Nationality	Zip

Generalization Lattice

Nationality	Zip
*	1306*
*	1305*
*	1485*



Nationality	Zip
American	130**
Japanese	130**
Japanese	148**

Nationality	Zip
American	1306*
Japanese	1305*
Japanese	1485*

→ **Suppress strictly more information**

Monotonic functions allow efficient lattice searches.

Theorem: If T satisfies L-Diversity, then any further generalization T* also satisfies L-Diversity.

- Analogous monotonicity properties have been exploited to build efficient algorithms for k-Anonymity.
  - Incognito
  - Mondrian
  - Hilbert

Anatomy: Bucketization Algorithm

[Xiao, Tao SIGMOD 2007]

non-sensitive				sensitive
Name	Zip	Age	Sex	Disease
Bob	14850	23	M	Flu
Charlie	14850	24	M	Flu
Dave	14850	25	M	Lung Cancer
Ed	14850	27	M	Lung Cancer
Frank	14853	29	M	Mumps
Gloria	14850	21	F	Flu
Hannah	14850	22	F	Flu
Irma	14853	24	F	Breast Cancer
Jessica	14853	26	F	Ovarian Cancer
Karen	14853	28	F	Heart Disease

non-sensitive				sensitive
Name	Zip	Age	Sex	Disease
*	1485*	2*	M	Flu Lung Cancer Mumps Flu Lung Cancer
*	1485*	2*	F	Flu Breast Cancer Flu Heart Disease Ovarian Cancer

Figure 2. 5-anonymous table

non-sensitive			
Name	Zip	Age	Sex
Bob	14850	23	M
Charlie	14850	24	M
Dave	14850	25	M
Ed	14850	27	M
Frank	14853	29	M
Gloria	14850	21	F
Hannah	14850	22	F
Irma	14853	24	F
Jessica	14853	26	F
Karen	14853	28	F

sensitive
Disease
Flu
Lung Cancer
Mumps
Flu
Lung Cancer
Flu
Breast Cancer
Flu
Heart Disease
Ovarian Cancer

Figure 3. Bucketized table

L-Diversity: Summary

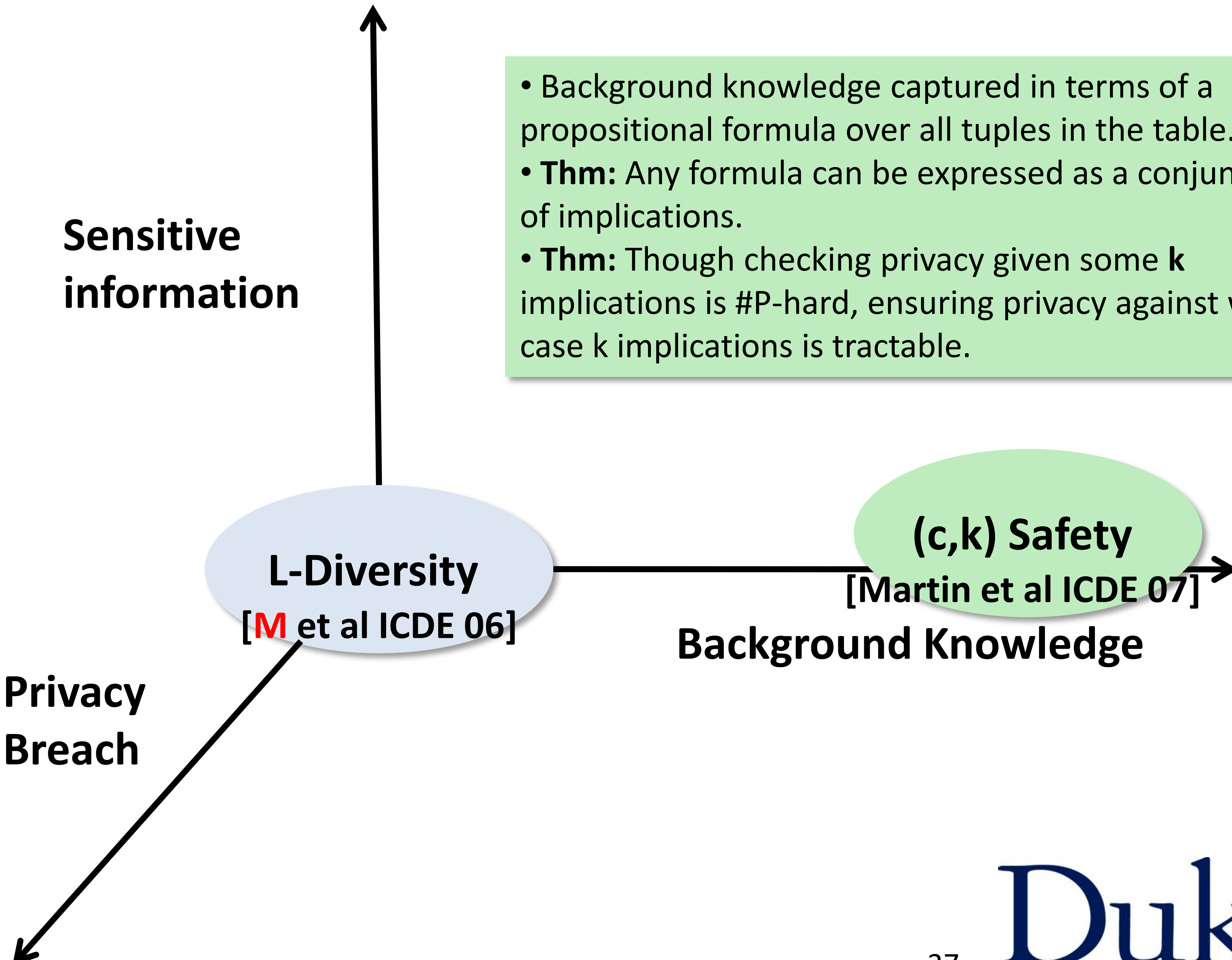
- Formally specified privacy model.

L-Diversity Principle:
Each group of tuples sharing the same Q-ID must have at least L distinct sensitive values that are roughly of equal proportions.

- Permits efficient and practical anonymization algorithms.

Duke UNIVERSITY

• Background knowledge captured in terms of a propositional formula over all tuples in the table.
• **Thm:** Any formula can be expressed as a conjunction of implications.
• **Thm:** Though checking privacy given some **k** implications is #P-hard, ensuring privacy against worst case k implications is tractable.



Duke UNIVERSITY

Background Knowledge

- Adversaries may possess more complex forms of background knowledge
  - If Alice has the flu, then her husband Bob very likely also has the flu.

- In general, background knowledge can be a boolean expression over individuals and their attribute values.
  - $t_{Ed}[disease] \neq flu$
  - $t_{Alice}[disease] = Flu \rightarrow t_{Bob}[disease] = Flu$
  - $(t_{Alice}[disease] = flu \vee t_{Alice}[disease] = cancer)$
    $\wedge (t_{Bob}[disease] = flu \vee t_{Bob}[disease] = cancer)$

Background Knowledge

Theorem: Any boolean expression can be written as a conjunction of *basic implications* of the form:

$$(\wedge_{i \in [m]} A_i) \rightarrow (\vee_{j \in [n]} B_j)$$

Disclosure Risk

non-sensitive				sensitive
Name	**Zip**	**Age**	**Sex**	**Disease**
Bob	14850	23	M	Flu
Charlie	14850	24	M	Lung Cancer
Dave	14850	25	M	Mumps
Ed	14850	27	M	Flu
Frank	14853	29	M	Lung Cancer
Gloria	14850	21	F	Flu
Hannah	14850	22	F	Breast Cancer
Irma	14853	24	F	Flu
Jessica	14853	26	F	Heart Disease
Karen	14853	28	F	Ovarian Cancer

- Suppose you publish bucketization *T**,

$$disclosure = max_{t \epsilon T, s \epsilon S, \phi} P[t[S] = s \mid T^* \wedge \phi]$$

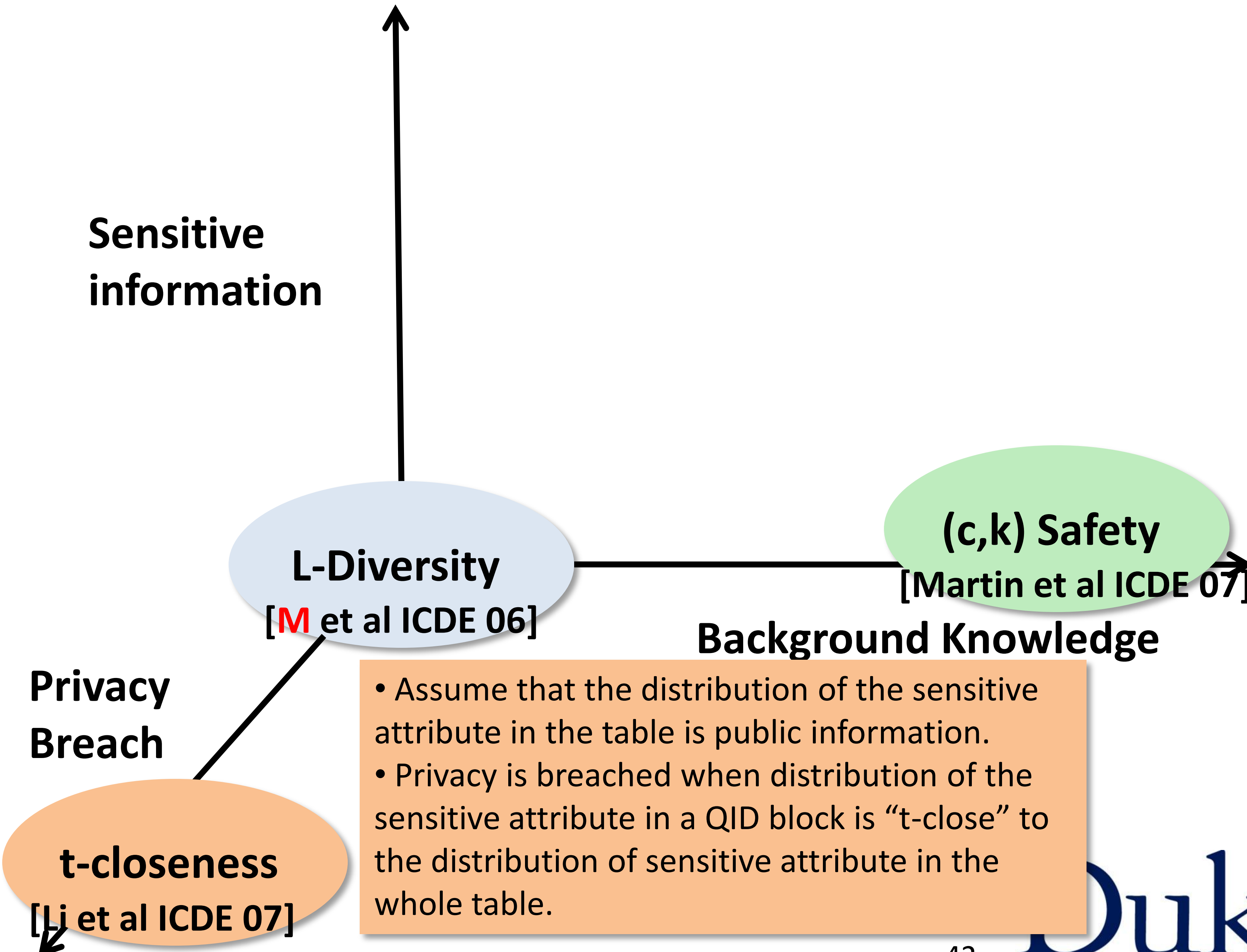
where, ϕ ranges over all boolean expressions which can be expressed as a conjunction of at most k basic implications.

Efficiently computing disclosure risk

- Disclosure is maximized when each implication is simple.

$$\wedge_{i \in [k]} (A_i \rightarrow B)$$

- Max disclosure can be computed in poly time (using dynamic programming)



- Assume that the distribution of the sensitive attribute in the table is public information.
- Privacy is breached when distribution of the sensitive attribute in a QID block is “t-close” to the distribution of sensitive attribute in the whole table.

Bounding posterior probability alone may not provide privacy

- Bob:
  - 52 years old
  - Earns 11K
  - Lives in 47909

- Suppose adversary knows distribution of disease in the entire table.
  - Pr[Bob has Flu] = 1/9

Disease
gastric ulcer
gastritis
stomach cancer
gastritis
flu
bronchitis
bronchitis
pneumonia
stomach cancer

Duke UNIVERSITY

Bounding posterior probability alone may not provide privacy

- Bob:
  - 52 years old
  - Earns 11K
  - Lives in 47909

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

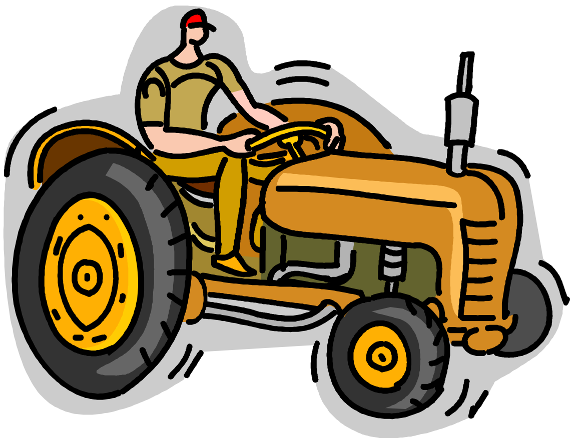
- After 3-diverse table is published.
  - Pr[Bob has Flu] = 1/3
- 1/9 → 1/3 is a large jump in probability

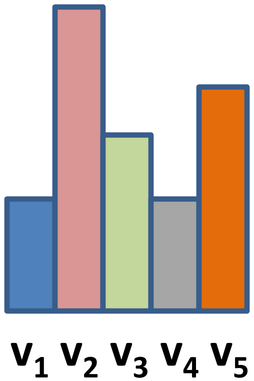
T-closeness principle

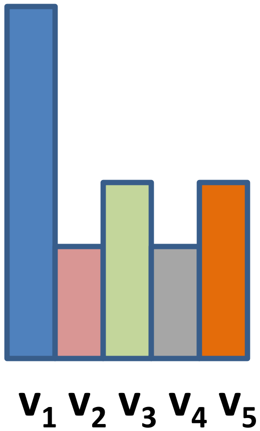
Distribution of sensitive attribute within each equivalence class should be "close" to the distribution of sensitive attribute in the entire table.

- Closeness is measured using Earth Mover's Distance.

Earth Mover's Distance

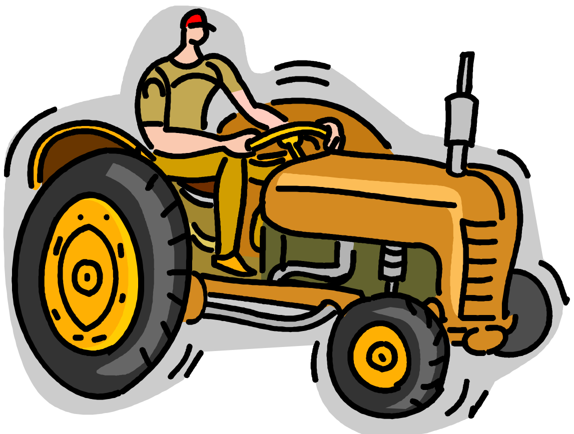


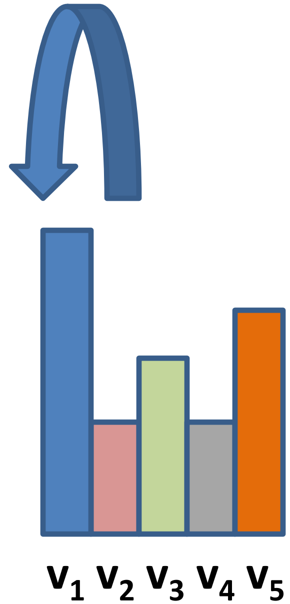


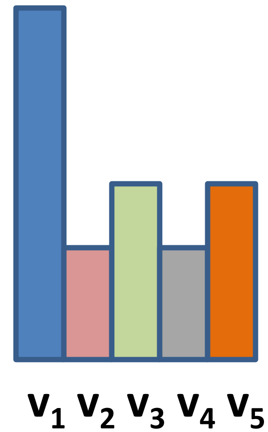


Duke UNIVERSITY

Earth Mover's Distance

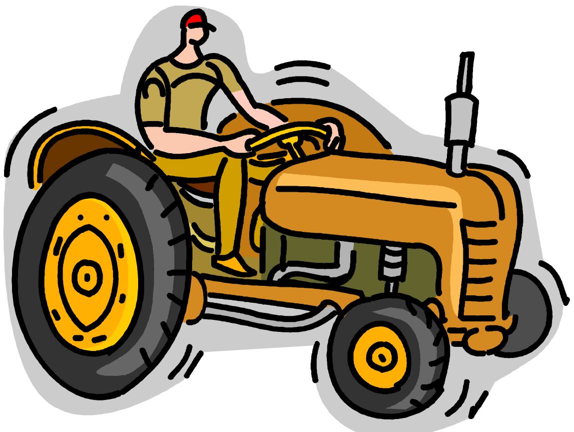


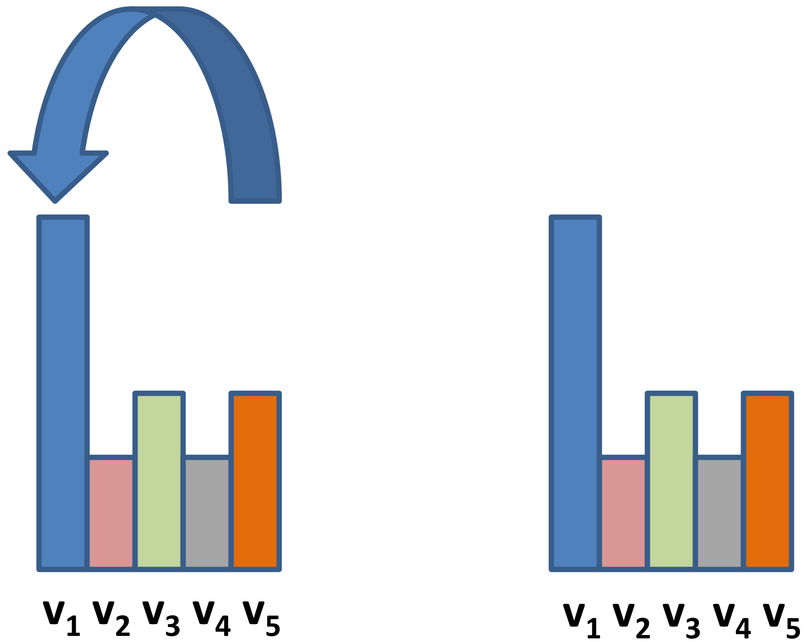




Distance = Cost of moving mass from v2 to v1 (f_{21})

Earth Mover's Distance





Distance = Cost of moving mass from v2 to v1 (f_{21})
+ cost of moving mass from v5 to v1 (f_{51})

If the values are numeric, cost can depend not only on amount of "earth" moved, but also the distance it is moved (d_{21} and d_{51}).

Earth Movers Distance

$$\mathrm{WORK}(P, Q, \mathbf{F}) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij},$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m,\ 1 \leq j \leq n \qquad (1)$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{\mathbf{p}_i} \qquad 1 \leq i \leq m \qquad (2)$$

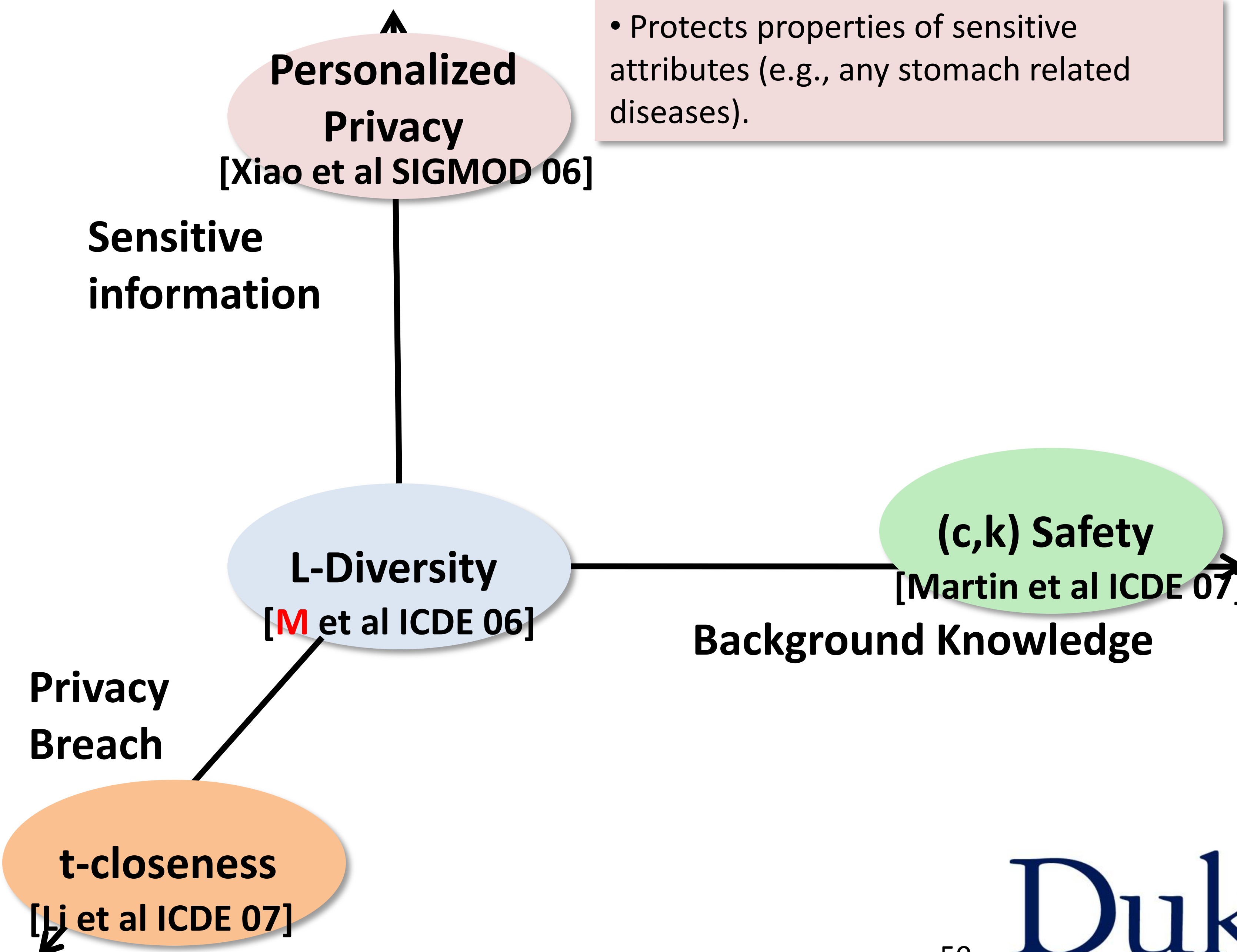
$$\sum_{i=1}^{m} f_{ij} \leq w_{\mathbf{q}_j} \qquad 1 \leq j \leq n \qquad (3)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{\mathbf{p}_i}, \sum_{j=1}^{n} w_{\mathbf{q}_j}\right), \qquad (4)$$

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}},$$

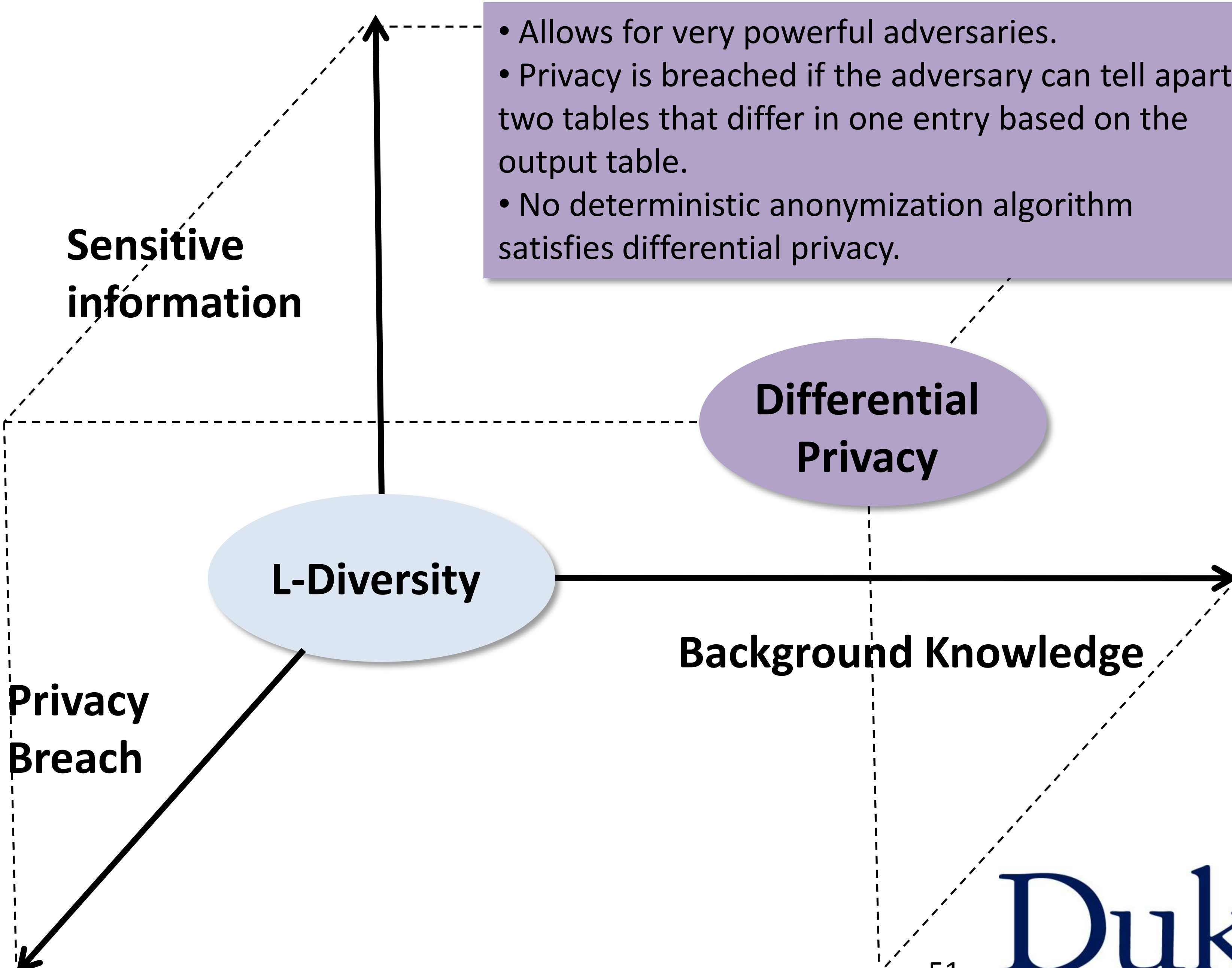
Original probability mass in the two distributions p and q which are being compared

- Protects properties of sensitive attributes (e.g., any stomach related diseases).



Duke UNIVERSITY

- Allows for very powerful adversaries.
- Privacy is breached if the adversary can tell apart two tables that differ in one entry based on the output table.
- No deterministic anonymization algorithm satisfies differential privacy.



Duke UNIVERSITY

Summary

- Adversaries can use background knowledge to learn sensitive information about individuals even from datasets that satisfy some measure of anonymity

- Many privacy definitions proposed for handling background knowledge
  - State of the art: Differential privacy (lecture 8)

- Next Class: Simulatability of algorithms

References

L. Sweeney, "*K-Anonymity: a model for protecting privacy*", IJUFKS 2002

A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, "*L-Diversity: Privacy beyond k-anonymity*", ICDE 2006

D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J. Halpern, "*Worst Case Background Knowledge*", ICDE 2007

N. Li, T. Li, S. Venkitasubramanian, "*T-closeness: privacy beyond k-anonymity and l-diversity*", ICDE 2007

X. Xiao & Y. Tao, "*Personalized Privacy Preservation*", SIGMOD 2006