Attacks on k-anonymous table

Table 1.2. Generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Heart disease
(Bruce)	2	20–29	Any	130**	Any	Heart disease
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
(Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

Attacks on k-anonymous table

Table 1.2. Generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Asthma
(Bruce)	2	20–29	Any	130**	Any	Asthma
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
(Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

Prior and posterior beliefs

Table 1.2. Generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Heart disease
(Bruce)	2	20–29	Any	130**	Any	Heart disease
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
(Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

Prior knowledge

$\alpha = p(t[S] = s \mid t[Q] = q)$

Posterior knowledge

$\beta = p(t[S] = s | t[Q] = q$
$\wedge \ \exists t^* \in T^*, t \to t^*)$

How to prevent *homogeneity* attacks?

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

ℓ-diversity

An equivalence class (E) is ℓ-diverse if it contains at least ℓ *well-represented* values for the sensitive attribute S.

A table is ℓ–diverse if every E is ℓ-diverse.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Distinct ℓ-diversity

At least ℓ distinct values for the sensitive attribute in each equivalence class.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	$*$	Heart Disease
4	1305*	≤ 40	$*$	Viral Infection
9	1305*	≤ 40	$*$	Cancer
10	1305*	≤ 40	$*$	Cancer
5	1485*	> 40	$*$	Cancer
6	1485*	> 40	$*$	Heart Disease
7	1485*	> 40	$*$	Viral Infection
8	1485*	> 40	$*$	Viral Infection
2	1306*	≤ 40	$*$	Heart Disease
3	1306*	≤ 40	$*$	Viral Infection
11	1306*	≤ 40	$*$	Cancer
12	1306*	≤ 40	$*$	Cancer

Skewness within equivalence classes

Age	Response (sensitive)
15	Y
21	N
19	Y
66	N
58	Y
71	N
….	…..
55	N

100,000 records

	Age	Response
Equivalence Class (n=1000)	15	N
	…	N
	19	Y
	…	…
	…	…
Equivalence Class (n=1000)	71	Y
	57	N
	55	Y
	18	Y

2-diverse version of the table

Entropy ℓ-diversity

The entropy of an equivalent class E is defined to be

$$\mathrm{H(E)} = -\sum_{s \in S} p(E,s) \log\big(p(E,s)\big)$$

An equivalence class E has entropy ℓ-diversity if

$$\mathrm{H(E)} \geq \log(l)$$

A table is ℓ-diverse if each E is ℓ-diverse.

Entropy ℓ-diversity

The entropy of an equivalent class E is defined to be

$$\mathrm{H(E)} = -\sum_{s \in S} p(E,s)\log\big(p(E,s)\big)$$

An equivalence class E has entropy ℓ-diversity if

$$\mathrm{H(E)} \geq \log(l)$$

A table is ℓ-diverse if each E is ℓ-diverse.

Constraint: the full table must have $\boldsymbol{H(T) \geq \log(l)}$
because $\boldsymbol{H(T) \geq min(H(E_a), H(E_b))}$

Skewness in the population

Age	Response
15	N
21	N
19	Y
66	N
58	N
71	N
….	…..
55	N

100,000 Records, 99% negative

Age	Response
15	N
…	N
19	Y
…	…
…	…
71	N
57	Y
….	….
55	N
18	N

10,000 records

10,000 records

2-diverse version of the table

Recursive (c,ℓ)-diversity

Let m be the number of values in an equivalence class, and r_i $(1 \leq i \leq m)$ be the number of times that the *ith* most frequent sensitive value appears in an equivalence class **E**.

Then, E has recursive (c, ℓ)-diversity if $r_1 < c(r_l, r_{l+1} + \ ... + r_m)$

A table is said to have recursive (c, ℓ)-diversity if all of its equivalence classes have recursive (c, ℓ)-diversity.

Recursive (c,ℓ)-diversity

Table 1.2. Generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Heart disease
(Bruce)	2	20–29	Any	130**	Any	Heart disease
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
(Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

m=3

r1 = 2

r2= 1 = r3

Then, E has recursive (c, ℓ)-diversity

if $r_1 < c(r_2 + r_3)$

Recursive (c,ℓ)-diversity

$$r_1 < c(r_l, r_{l+1} + \ \ldots + r_m)$$

The count of the most frequent item must be *less than* a constant multiple of the total count of the least (m-l) frequent items.

Limitations of ℓ-diversity: drastic loss of utility

Zip	Disease
15152	Y
15163	Y
15784	Y
18784	N
18654	N
18744	N

Zip	Disease
15152	Y
18784	N
15163	Y
18784	N
15784	Y
18744	N

zip	Disease
1****	Y
1****	N
1****	Y
1****	N
1****	Y
1****	N

Sometimes ℓ-diversity may be unnecessary

Age	Response
15	N
21	N
19	Y
66	N
58	N
71	N
….	…..
55	N

100,000 Records, 99% negative

	Age	Response
10,000 records	15	N
	…	N
	19	Y
	…	…
	…	…
10,000 records	71	N
	57	Y
	….	….
	55	N
	18	N

2-diverse version of the table

Limitations of ℓ-diversity: wrong inference because of skewness

Age	Disease
15	N
21	N
19	Y
66	N
58	N
71	N
….	…..
55	Y

100,000 Records, 99% negative

Equivalence class

Age	Disease
15	N
21	N
19	Y
55	Y
….	….
…	…
…	…

2-diverse version of the table

Limitations of ℓ-diversity: similarity attack

Age	Disease
15	Lung cancer
21	Stomach cancer
19	No issue
66	Stomach cancer
58	Liver cancer
71	No issue
….	…..
55	Headache

Age	Response
15	Lung cancer
…	Stomach cancer
19	Liver cancer
…	…
…	…
71	…
….	…
55	…
18	…

Intuition behind the limitations of l-diversity

Intuitively, distributions that have the same level of (syntactic) diversity may provide very different levels of privacy, because

1. there are semantic relationships among the attribute value
2. different values have very different levels of sensitivity, e.g., headache can have vastly different level of sensitivity than cancer/diabetic
3. different distributions at the equivalent class level and population level can have adverse effects

Intuition behind t-closeness

Table 1.2. Generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Heart disease
(Bruce)	2	20–29	Any	130**	Any	Heart disease
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
(Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

Table 1.2. Generalized medical record table.

	Age	Gender	Zip Code	Nationality	Condition
(Ann)					Heart disease
(Bruce)					Heart disease
(Cary)					Viral infection
(Dick)					Viral Infection
(Eshwar)					Cancer
(Fox)					Flu
(Gary)					Heart disease
(Helen)					Flu
(Igor)					Cancer
(Jean)					Cancer
(Ken)					Cancer
(Lewis)					Cancer

Q

Prior belief about an individual, α

Changed belief about an individual, β

Intuition behind t-closeness

Table 1.2. Generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Heart disease
(Bruce)	2	20–29	Any	130**	Any	Heart disease
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
(Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

P

Prior belief about an individual, α

Changed belief after seeing distribution at the population level, β

Changed belief after looking at the (anonymized) table, γ

l −diversity minimizes $\gamma - \alpha$

t −closeness minimizes $\gamma - \beta$

t −closeness minimizes $\gamma - \beta$

Table 1.2. Generalized medical record table. Q

	Age	Gender	Zip Code	Nationality	Condition
(Ann)					Heart disease
(Bruce)					Heart disease
(Cary)					Viral infection
(Dick)					Viral Infection
(Eshwar)					Cancer
(Fox)					Flu
(Gary)					Heart disease
(Helen)					Flu
(Igor)					Cancer
(Jean)					Cancer
(Ken)					Cancer
(Lewis)					Cancer

Table 1.2. Generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Heart disease
(Bruce)	2	20–29	Any	130**	Any	Heart disease
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
P (Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

Q(heart disease) = 1/4

Q(Flu) = 1/6

P(heart disease) = 1/4

P(Flu) = 1/2

t-closeness

An equivalence class has t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t.

$$dist(P, Q) \leq t$$

A table is said to have t-closeness if all equivalence classes have t-closeness.

t-closeness

$$dist(P, Q) \leq t$$

Distance measures

1. Kulback-Liebler divergence
2. Jensen-Shannon
3. Earth mover’s distance

resources

Diversity: Privacy Beyond k-Anonymity

t-closeness: **Privacy beyond k-anonymity** and l-**diversity**