t-Closeness: Privacy beyond k-Anonymity and l-Diversity

paper by Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian

presentation by Caitlin Lustig for CS 295D

Definitions

- *Attributes*: explicit identifiers, ie SSN, address, name.
- *Quasi-identifiers*: non-explicit identifies, ie zip code, birthdate, gender.
- *Sensitive attributes*: attributes that should be private, ie disease and salary.
- *Identity disclosure*: when an individual is linked to a record.
- *Attribute disclosure*: new information about some individuals are revealed. Identity disclosure generally leads to attribute disclosure.
- *Equivalence class*: a set of records with the same anonymized data.

Problem Space

- Pre-existing privacy measures *k*-anonymity and *l*-diversity have flaws.
- *k-anonymity*-each equivalence class has at least k records to protect against identity disclosure.
- *k*-anonymity is vulnerable to *homogeneity attacks* and *background knowledge attacks*.
- *l-diversity*: distribution of a sensitive attribute in each equivalence class has at least *l* "well represented" values to protect against attribute disclosure.
- *l*-diversity is vulnerable to *skewness attacks* and *similarity attacks*.

k-anonymity

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table 1. Original Patients Table

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Table 2. A 3-Anonymous Version of Table 1

Definition:

k-anonymity-each equivalence class has at least k records to protect against identity disclosure.

Attacks on *k*-anonymity:

homogeneity attack: Bob is a 27-year old man living in zip code 47678 and Bob's record is in the table. So Bob corresponds to one of the first three records and must have heart disease.

background knowledge attack: Carl is a 32-year old man living in zip code 47622. Therefore he is in the last equivalence class in Table 2. If you know that Carl has a low risk for heart disease then you can conclude that Carl probably has cancer.

l-diversity

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

Definition:

l-diversity: distribution of a sensitive attribute in each equivalence class has at least *l* "well represented" values to protect against attribute disclosure.

Attacks on *l*-diversity:

Similarity Attack: Table 4 anonymizes table 3. It's sensitive attributes are *Salary* and *Disease*.

If you know Bob has a low salary (3k-5k) then you know that he has a stomach related disease.

This is because *l*-diversity takes into account the diversity of sensitive values in the group, but does not take into account the semantical closeness of the values.

l-diversity

	Zip Code	Age	Salary	Disease
1	476**	2*	3k	negative
2	476**	2*	4k	negative
3	476**	2*	5k	negative
4	476**	2*	6k	negative
5	4790*	>=40	7k	negative
6	4790*	>=40	8k	positive
7	4790*	>=40	9k	negative
8	4790*	>=40	10k	positive
9	476**	3*	11k	positive
10	476**	3*	12k	positive
11	476**	3*	13k	positive
12	476**	3*	14k	negative
13	4770*	4*	15k	negative
...	...	...	...	...
10,000	488**	>=60	16k	negative

We have 10,000 records about a virus that affects only 1% of the population.

For equivalence class 1, strong privacy measures probably aren't necessary because people don't have the disease don't care if their identity is discovered.

Skewness attack: The second equivalence class has an equal number of positive and negative records. This gives everyone in this equivalence class a 50% chance of having the virus, which is much higher than the real distribution. The third equivalence class has an even higher privacy risk.

l-diversity assumes that adversaries don't have access to the global distribution of sensitive attributes, however adversaries can learn the distribution by just looking at the table!

t-closeness overview

	Zip Code	Age	Salary	Disease
1	476**	2*	3k	negative
2	476**	2*	4k	negative
3	476**	2*	5k	negative
4	476**	2*	6k	negative
5	4790*	>=40	7k	negative
6	4790*	>=40	8k	positive
7	4790*	>=40	9k	negative
8	4790*	>=40	10k	positive
9	476**	3*	11k	positive
10	476**	3*	12k	positive
11	476**	3*	13k	positive
12	476**	3*	14k	negative
13	4770*	4*	15k	negative
...	...	...	...	...
10,000	488**	>=60	16k	negative

Privacy is measured by the information gain of an observer. The gain is the difference between the *prior belief* and the *posterior belief*. Each belief is denoted by B_n where *n* is the number of the belief.

B_0: Alice believes that Bob has the virus because he has been acting sick.

B_1: Alice gets a summary report of the table and learns that only 1% of the population has the virus. This distribution is *Q*, the distribution of the sensitive attribute in the whole table. She believes that Bob is in that one percent.

B_2: Alice takes a look at the table, and finds that Bob is in equivalence class 3 because he is 32 and lives in zip code 47623. She learns *P*, the distribution of the sensitive attribute values in this class. Based on *P* she decides that it is actually quite likely that Bob has the virus.

t-closeness overview

- *l*-diversity limits the gain between B_0 (belief before any knowledge of the table) and B_2 (belief after examining the table and the relevant equivalence class) by requiring that *P* (distribution in the equivalence class) has diversity.
- *Q* (global distribution in the table) should be treated as public information.
- If the change from B_0 to B_1 is large, means that the *Q* contains lots of new information. But we can't control people's access to *Q*, so we shouldn't worry about it.
- Therefore should focusing on limiting the gain between B_1 and B_2. We can do so by limiting the difference between *P* and *Q*. The closer *P* and *Q* are, the closer B_1 and B_2 are.

t-closeness definition

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

Distance measurements

Now that we've confirmed that limiting the difference between *P* and *Q* is the key to privacy, we need a way to measure the distance. Here are some naive measurements:

Variational distance:

$$\mathsf{D}[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2}|p_i - q_i|.$$

Kullback-Leibler (KL) distance:

$$\mathsf{D}[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} = H(\mathbf{P}) - H(\mathbf{P},\mathbf{Q})$$

Observe the problem with these measurements:

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

Q = {3k,4k,5k,6k,7k,8k,9k,10k,11k}
P_1 = {3k,4k,5k}
P_2 = {6k,8k,11k}

P_1 has more information leakage than P_2 because there are fewer people in that salary range and thus they are easier to identify, thus we should have $D[P_1,Q] > D[P_2,Q]$.

However, these algorithms just view 3k and 6k as different points and don't attach semantic meaning to them. They would calculate this wrong.

Earth Mover Distance

Definition:

The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. Intuitively, one distribution is seen as a mass of earth spread in the space and the other as a collection of holes in the same space. EMD measures the least amount of work needed to fill the holes with earth. A unit of work corresponds to moving a unit of earth by a unit of ground distance.

Earth Mover Distance

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

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq m \qquad (c1)$$

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i \qquad 1 \leq i \leq m \qquad (c2)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1 \qquad (c3)$$

Properties of EMD:

- $d_{i,j}$ is the ground distance between *i* in *P* and *j* in *Q*.
- $f_{i,j}$ is the flow of mass to transform *i* in *P* into *j* in *Q* using the minimal amount of work.
- *F* is the mass flow to transform *P* into *Q*.
- WORK(*P*,*Q*,*F*) = D[*P*,*Q*]. WORK is the work to transform P into Q.
- D[P,Q] is between 0 and 1.
- For any P_1 *and* P_2, D[P,Q]<=max(D[P_1,Q],D[P_2, Q]).

Earth Mover Distance

EMD gives us a method for determining the distance between two distributions but doesn't tell us how to determine the distance between two elements in the distributions. The way to do that will differ depending on the type of data we're using...

Numerical Distances

Ordered Distance: Distance between 2 *numerical attributes* (ie age) is based on the number of values between them in the total order.

Formally, let $r_i = p_i - q_i$,(i=1,2,...,m), then the distance between $\mathbf{P}$ and $\mathbf{Q}$ can be calculated as:

$$D[\mathbf{P}, \mathbf{Q}] = \frac{1}{m-1}(|r_1| + |r_1 + r_2| + ... + |r_1 + r_2 + ... r_{m-1}|)$$

$$= \frac{1}{m-1} \sum_{i=1}^{i=m} \left| \sum_{j=1}^{j=i} r_j \right|$$

Minimal work can be achieved by satisfying all elements of *Q* sequentially

Equal Distance

For *categorical attributes* (ie diseases), order does not always matter. We can either view the ground distance between 2 categorical attributes as always being 1 (equal distance). "As the distance between any two values is 1, for each point that $p_i - q_i > 0$, one just needs to move the extra to some other points."

Equal distance:

$$D[\mathbf{P}, \mathbf{Q}] = \frac{1}{2}\sum_{i=1}^{m}|p_i - q_i| = \sum_{p_i \geq q_i}(p_i - q_i) = -\sum_{p_i < q_i}(p_i - q_i)$$

Hierarchical Distance

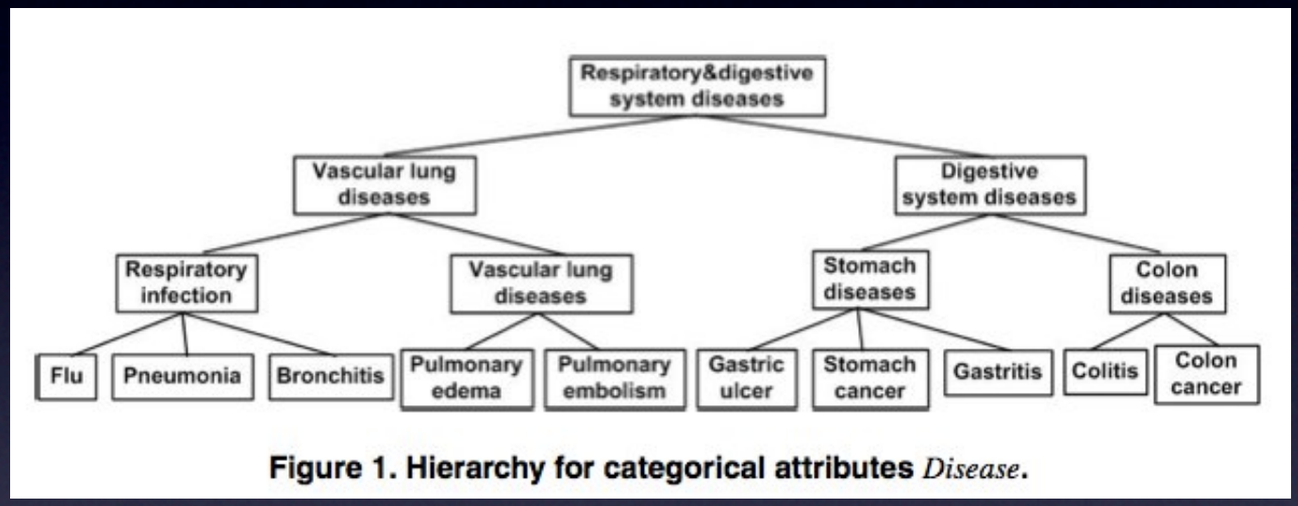


Figure 1. Hierarchy for categorical attributes *Disease*.

Another way to measure categorical attributes is taking into account the hierarchical distance.

Hierarchical distance: *H* is the height of the domain hierarchy. The distance between two leaves of the hierarchy is defined to be $level(v_1,v_2)/H$ where $level(v_1,v_2)$ is the height of the lowest common ancestor node of v_1 and v_1.

Hierarchical Distance

$$extra(N) = \begin{cases} p_i - q_i & \text{if N is a leaf} \\ \sum_{C \in Child(N)} extra(C) & \text{otherwise} \end{cases}$$

Child(N) is the set of all leaf nodes below *N*.

$$pos\_extra(N) = \sum_{C \in Child(N) \wedge extra(C) > 0} |extra(C)|$$

$$neg\_extra(N) = \sum_{C \in Child(N) \wedge extra(C) < 0} |extra(C)|$$

The *extra* function has the property that the sum of *extra* values for nodes at the same level is 0.

$$cost(N) = \frac{height(N)}{H} \min(pos\_extra(N), neg\_extra(N))$$

Therefore:

$$\mathrm{D}[\mathbf{P}, \mathbf{Q}] = \sum_N cost(N)$$

Properties of *t*-closeness

Now we have a framework for calculating EMD and thus achieving *t*-closeness.
What are the properties of *t*-closeness?

- *Generalization property:* If *A* and *B* are generalizations on the table *T* such that *A* is more general than *B* and *T* satisfies *t*-closeness using *B*, then *T* also satisfies *t*-closeness using *A*.

- *Subset Property:* If *C* is a set of attributes in the table *T* and if *T* satisfies *t*-closeness with respect to *C*, then *T* also satisfies *t*-closeness with respect to any set of attributes *D* such that *D* is a subset of *C*.

Example of *t*-closeness

Remember this slide? Now let's calculate the EMD and create a *t*-close table.

Observe the problem with these measurements:

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

Q = {3k,4k,5k,6k,7k,8k,9k,10k,11k}
P_1 = {3k,4k,5k}
P_2 = {6k,8k,11k}

P_1 has more information leakage than P_2 because there are fewer people in that salary range and thus they are easier to identify, thus we should have $D[P_1,Q]>D[P_2,Q]$.

However, these algorithms just view 3k and 6k as different points and don't attach semantic meaning to them. They would calculate this wrong.

Example of *t*-closeness

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

P_1 = {3k, 4k, 5k}

Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

One optimal mass flow that transforms P_1 to Q is to move 1/9 probability mass across the following pairs: (5k→11k), (5k→10k), (5k→9k), (4k→8k), (4k→7k), (4k→6k), (3k→5k), (3k→4k). The cost of this is 1/9× (6+5+4+4+3+2+2+1)/8 = 27/72 = 3/8 = 0.375.

Remember: for numerical attributes, minimal work can be achieved by satisfying all elements of Q sequentially

Example of *t*-closeness

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

P_2 = {6k, 8k, 11k}

Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

One optimal mass flow that transforms P_1 to Q is to move 1/9 probability mass across the following pairs: (11k→10k), (11k→9k), (8k→8k), (8k→7k), (8k→6k), (6k→5k), (6k→4k), (6k→3k). The cost of this is 1/9× (1+2+0+1+2+1+2+3)/8 = 12/72 = 3/18 = 0.167.

Example of *t*-closeness

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

P_1 = {3k, 4k, 5k}

P_2 = {6k, 8k, 11k}

Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

D[P_1,Q] is 0.375 and D[P_2,Q] has a distance of 0.167. Therefore, P_2 reveals less private data.

Example of *t*-closeness

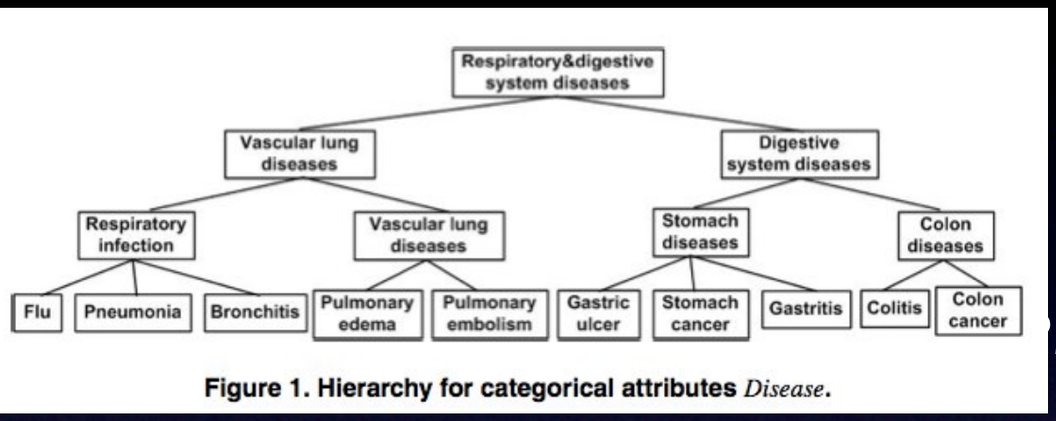


Figure 1. Hierarchy for categorical attributes *Disease*.

P_1 = {gastric ulcer, gastritis, stomach cancer}

P_2 = {gastritis, flu, bronchitis}

Q = {gastric ulcer, gastritis, stomach cancer, gastritis, flu, bronchitis, bronchitis, pneumonia, stomach cancer}

Difference between gastric ulcer and gastritis is 1/3.
Difference between gastric ulcer and colitis is 2/3.
Difference between gastric ulcer and flu is 3/3.

Remember: The distance between two leaves of the hierarchy is defined to be $level(v_1,v_2)/H$ where $level(v_1,v_2)$ is the height of the lowest common ancestor node of v_1 and v_1.

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

Example of *t*-closeness

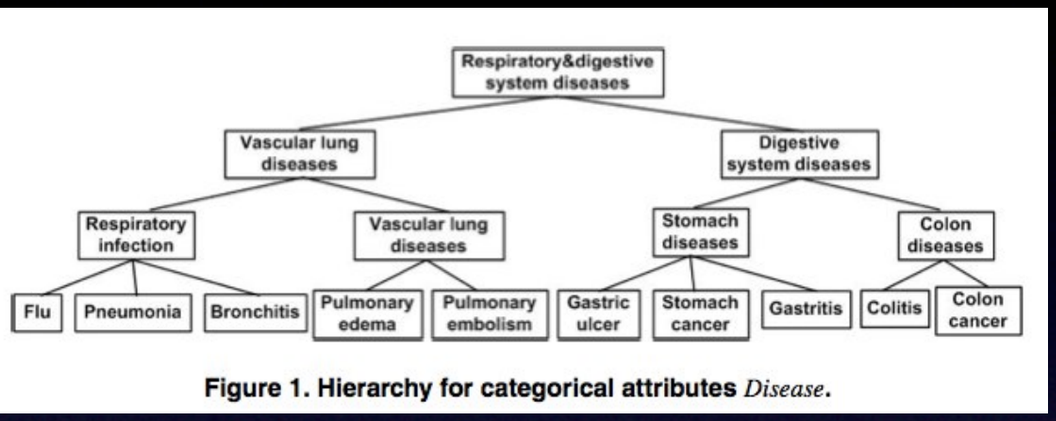


Figure 1. Hierarchy for categorical attributes *Disease*.

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

P_1 = {gastric ulcer, gastritis, stomach cancer}

P_2 = {gastritis, flu, bronchitis}

Q = {gastric ulcer, gastritis, stomach cancer, gastritis, flu, bronchitis, bronchitis, pneumonia, stomach cancer}

D[P_1,Q] is 0.5 and D[P_2,Q] has a distance of 0.278. Therefore, P_2 reveals less private data.

Example of *t*-closeness

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	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

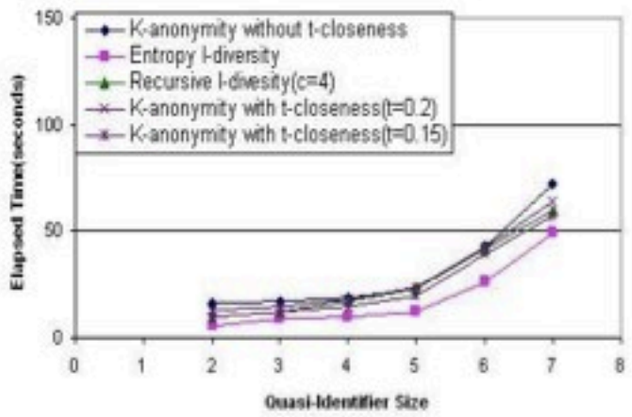
	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

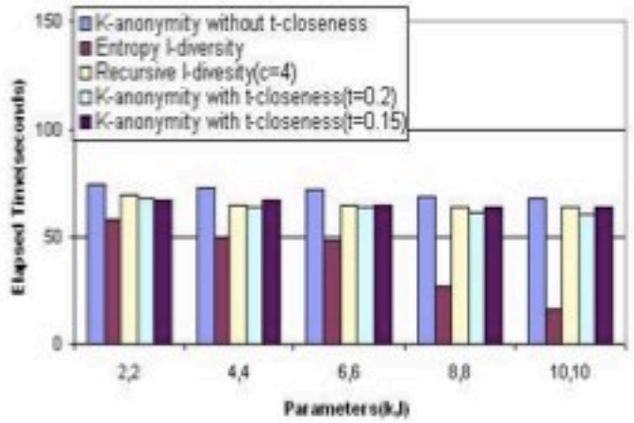
Experiment

- Used the adult dataset from the UCI machine learning dataset--data collected from the US census. Used 30162 records.
- Used 9 attributes from the dataset--age, workclass, education, country, marital status, race, gender, occupation, and salary. Occupation and salary are the sensitive attributes.
- Compared *k*-anonymity, entropy *l*-diversity, recursive *(c,l)* diversity, *k*-anonymity with *t*-closeness(t=0.2), and *k*-anonymity with *t*-closeness(*t*=0.15).
- Results: *l*-diversity runs faster than the other four measures. Entropy *l*-diversity has the worst data quality. *k*-anonymity has slightly better data quality than *t*-closeness.

Experimental Results

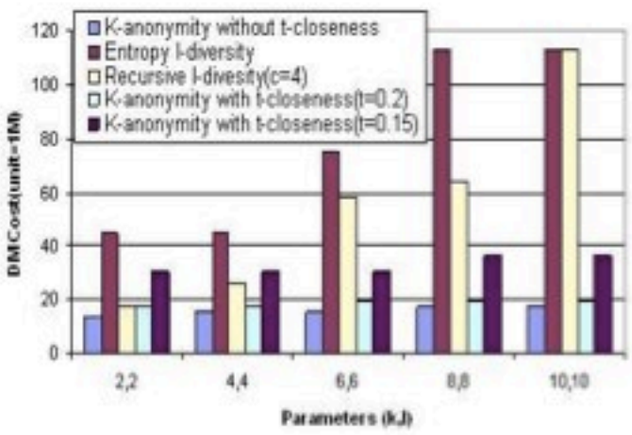


(a) Varied QI size for $k = 5, l = 5$

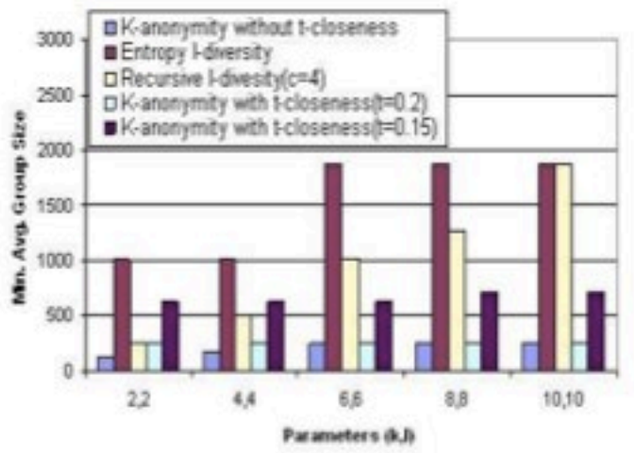


(b) Varied parameters k and l

Figure 3. Efficiency of the Five Privacy Measures.



(a) Discernibility metric cost



(b) Minimal average group size

Figure 4. Data Quality of the Five Measures.