L-Diversity: Privacy Beyond K-Anonymity

Proposed by-

Ashwin Machanavajjhala,

Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam

Presented by- Akash Yadav

Overview

- Introduction
- Attacks on k-Anonymity
- Bayes Optimal Privacy
- l-Diversity Principle
- l-Diversity Instantiations
- Multiple Sensitive Attributes
- Monotonicity Property
- Utility
- Conclusion

Background

- Large amount of person-specific data has been collected in recent years
  - Both by governments and by private entities
- Data and knowledge extracted by data mining techniques represent a key asset to the society
  - Analyzing trends and patterns
  - Formulating public policies
- Laws and regulations require that some collected data must be made public
  - For example, Census data

What About Privacy?

- First thought: anonymize the data
- How?
- Remove "personally identifying information" (PII)
  - Name, Social Security number, phone number, email, address…
  - Anything that identifies the person directly
- Is this enough?

Re-identification by Linking

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Classification of Attributes

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized dataset with other datasets

- Sensitive attributes
  - Medical records, salaries, etc.
  - These attributes is what the researchers need, so they are always released directly

Key Attribute		Quasi-identifier			Sensitive attribute
Name		DOB	Gender	Zipcode	Disease
Andre		1/21/76	Male	53715	Heart Disease
Beth		4/13/86	Female	53715	Hepatitis
Carol		2/28/76	Male	53703	Brochitis
Dan		1/21/76	Male	53703	Broken Arm
Ellen		4/13/86	Female	53706	Flu
Eric		2/28/76	Female	53706	Hang Nail

K-Anonymity

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

Attacks on K-anonymity

- Homogeneity Attacks
- Background Knowledge Attacks

Homogeneity Attacks

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Original Table

	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

4-anonymous Table

Since Alice is Bob's neighbor, she knows that Bob is a 31-year-old American male who lives in the zip code 13053. Therefore, Alice knows that Bob's record number is 9,10,11, or 12. She can also see from the data that Bob has cancer.

Background Knowledge Attacks

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Original Table

	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

4-anonymous Table

Alice knows that Umeko is a 21 year-old Japanese female who currently lives in zip code 13068. Based on this information, Alice learns that Umeko's information is contained in record number 1,2,3, or 4. With additional information, Umeko being Japanese and Alice knowing that Japanese have an extremely low incidence of heart disease, Alice can concluded with near certainty that Umeko has a viral infection.

Weaknesses in k-anonymous tables

Given these two weaknesses there needs to be a stronger method to ensure privacy.

Based on this, the authors begin to build their solution.

Adversaries Background Knowledge

- The adversary has access to T* and knows it was derived from table T. The domain of each attribute is also known.
- The adversary may also have instance level background knowledge.
- The adversary may also know demographic background data such as the probability of a condition given an age.

Bayes-Optimal Privacy

- Models background knowledge as a probability distribution over the attributes and uses Bayesian inference techniques to reason about privacy.
- However, Bayes-Optimal Privacy is only used as a starting point for a definition of privacy so there are 2 simplifying assumptions made.
  - T is a simple random sample of a larger population.
  - Assume a single sensitive value

Prior belief is defined as:

Alice's *prior belief*, $\alpha_{(q,s)}$, that Bob's sensitive attribute is s given that his nonsensitive attribute is q, is just her background knowledge:

$$\alpha_{(q,s)} = P_f\left(t[S] = s \middle| t[Q] = q\right)$$

Posterior belief is defined as:

After Alice observes the table T^*, her belief about Bob's sensitive attribute changes. This new belief, $\beta_{(q,s,T^*)}$, is her *posterior belief*:

$$\beta_{(q,s,T^*)} = P_f\left(t[S] = s \middle| t[Q] = q \wedge \exists t^* \in T^*, t \xrightarrow{*} t^*\right)$$

Prior belief and posterior belief are used are used to gauge the attacker's success.

Calculating the posterior belief

Theorem 3.1 *Let q be a value of the nonsensitive attribute Q in the base table T; let q^* be the generalized value of q in the published table T^*; let s be a possible value of the sensitive attribute; let $n_{(q^*,s')}$ be the number of tuples $t^* \in T^*$ where $t^*[Q] = q^*$ and $t^*[S] = s'$; and let $f(s' \mid q^*)$ be the conditional probability of the sensitive attribute conditioned on the fact that the nonsensitive attribute Q can be generalized to q^*. Then the following relationship holds:*

$$\beta_{(q,s,T^*)} = \frac{n_{(q^*,s)} \frac{f(s|q)}{f(s|q^*)}}{\sum_{s' \in S} n_{(q^*,s')} \frac{f(s'|q)}{f(s'|q^*)}} \qquad (1)$$

Privacy Principles

Positive Disclosure: Publishing the table T $\star$ that was derived from T results in a positive disclosure if the adversary can correctly identify the value of a sensitive attribute with high probability.

Negative disclosure: Publishing the table T $\star$ that was derived from T results in a negative disclosure if the adversary can correctly eliminate some possible values of the sensitive attribute (with high probability)

Principle 1 (Uninformative Principle) *The published table should provide the adversary with little additional information beyond the background knowledge. In other words, there should not be a large difference between the prior and posterior beliefs.*

Drawbacks to Bayes-Optimal Privacy

- Insufficient knowledge because the publisher is unlikely to know the full distribution of sensitive and non-sensitive attributes over the full population.
- The data publisher does not know the knowledge of a would be attacker.
- Instance level knowledge cannot be modeled.
- There are likely to be many adversaries with varying levels of knowledge

L-Diversity Principle

Theorem 3.1 defines a method of calculating the observed belief of the adversary

In the case of positive disclosures, Alice wants to determine Bob's sensitive attribute with a very high probability. Given Theorem 3.1 this can only happen when:

$$\exists s,\ \forall s' \neq s, \quad n_{(q^*,s')}\frac{f(s'|q)}{f(s'|q^*)} \ll n_{(q^*,s)}\frac{f(s|q)}{f(s|q^*)} \qquad (2)$$

The condition of equation 2 can be satisfied by a lack of diversity in the sensitive attribute(s) and/or strong background knowledge.

Lack of diversity in the sensitive attribute can be described as follows:

$$\forall s' \neq s, \quad n_{(q^\star, s')} \ll n_{(q^\star, s)} \tag{3}$$

- Equation 3 indicates that almost all tuples have the same value as the sensitive value and therefore the posterior belief is almost 1.

- To ensure diversity and to guard against Equation 3 is to require that a q⋆-block has at least l ≥ 2 different sensitive values such that the l most frequent values (in the q⋆-block) have roughly the same frequency. We say that such a q⋆- block is *well-represented by l sensitive values*.

An attacker may still be able to use background knowledge when the following is true

$$\exists s', \quad \frac{f(s'|q)}{f(s'|q^\star)} \approx 0 \qquad (4)$$

This equation states that Bob with quasi-identifier t[Q] = q is much less likely to have sensitive value s´ than any other individual in the q⋆-block.

Revisiting the example

1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection

Suppose we consider an equivalence class for the example of background knowledge attack shown earlier.

Here Alice has background knowledge that Japanese people are less prone to heart disease.

$\therefore$ f (s´|q)=0 ($\because$ The probability that Umeko has heart disease

given her non sensitive attribute as 'Japanese' is 0).

Also, f(s´|q*)=2/4

$\therefore$ f(s´|q)/f(s´|q*)=0.

- In spite of such background knowledge, if there are l "well represented" sensitive values in a q⋆-block, then Alice needs l − 1 damaging pieces of background knowledge to eliminate l − 1 possible sensitive values and infer a positive disclosure!

L-Diversity Principle

Given the previous discussions, we arrive at the l-Diversity principle:

Principle 2 (ℓ-Diversity Principle) *A q^*-block is ℓ-diverse if contains at least ℓ "well-represented" values for the sensitive attribute S. A table is ℓ-diverse if every q^*-block is ℓ-diverse.*

Revisiting the example

	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

4-anonymous table

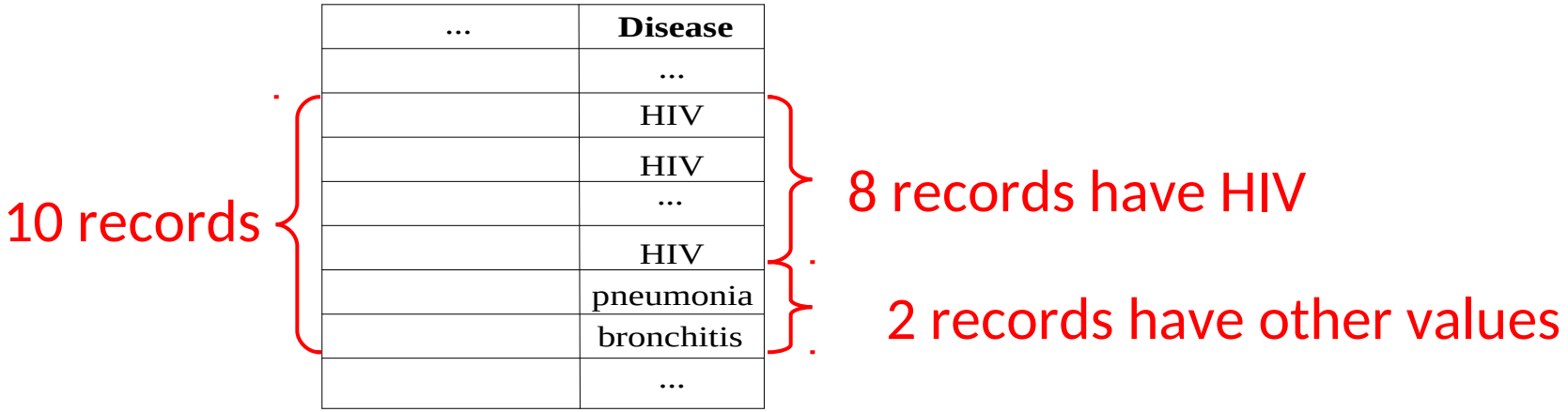
	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart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

3 diverse table

- Using a 3-diverse table, we no longer are able to tell if Bob (a 31 year old American from zip code 13053) has cancer.
- We also cannot tell if Umeko(a 21 year old Japanese from zip code 13068) has a viral infection or cancer.

Distinct l-Diversity

- Each equivalence class has at least l well-represented sensitive values
- Doesn't prevent probabilistic inference attacks



L-Diversity Instantiations

- Entropy l-Diversity
- Recursive (c, l) Diversity
- Positive Disclosure-Recursive (c, l)-Diversity
- Negative/Positive Disclosure-Recursive (c1, c2, l)-Diversity

Entropy l-Diversity

Definition 4.1 (Entropy ℓ-Diversity) *A table is* Entropy ℓ-Diverse *if for every q^*-block*

$$-\sum_{s \in S} p_{(q^*,s)} \log(p_{(q^*,s')}) \geq \log(\ell)$$

where $p_{(q^*,s)} = \frac{n_{(q^*,s)}}{\sum_{s' \in S} n_{(q^*,s')}}$ *is the fraction of tuples in the q^*-block with sensitive attribute value equal to s.*

- Here every q⋆-block has at least l distinct values for the sensitive attribute
- This implies that for a table to be entropy l-Diverse, the entropy of the entire table must be at least log(l).
- Therefore, entropy l-Diversity may be too restrictive to be practical.

Recursive (c, l) Diversity

- Less restrictive than entropy l-diversity
- Let s1, …, sm be the possible values of sensitive attribute S in a q*-block
- Assume, we sort the counts n(q*,s1), …, n(q*,sm) in descending order with the resulting sequence r1, …, rm.
- We can say a q*-block is recursive (c,l)-diverse if r1 < c(r2+ …. +rm) for a specified constant c.

Positive Disclosure-Recursive (c, l)-Diversity

Some cases of positive disclosure may be acceptable such as when medical condition is “healthy”.

To allow these values the authors define pd-recursive (c, l)-diversity

Negative/Positive Disclosure-Recursive (c1, c2, l)-Diversity

Npd-recursive (c1, c2, l)-diversity prevents negative disclosure by requiring attributes for which negative disclosure is not allowed to occur.

Multiple Sensitive Attributes

- Previous discussions only addressed single sensitive attributes.
- Suppose S and V are two sensitive attributes, and consider the q*-block with the following tuples:

  {(q ,s1,v1),(q ,s1,v2),(q ,s2,v3),(q ,s3,v3)}.

- This q*-block is 3-diverse (actually recursive (2,3)-diverse) with respect to S (ignoring V) and 3-diverse with respect to V (ignoring S). However, if we know that Bob is in this block and his value for S is not s1 then his value for attribute V cannot be v1 or v2, and therefore must be v3.
- To address this problem we can add the additional sensitive attributes to the quasi-identifier.

Implementing Privacy Preserving Data Publishing

- Domain generalization is used to define a generalization lattice.
- For discussion, all non-sensitive attributes are combined into a multi-dimensional attribute (Q) where the bottom element on the lattice is the domain of Q and the top of the lattice is the domain where each dimension of Q is generalized to a single value.

Implementing Privacy Data Publishing (cont.)

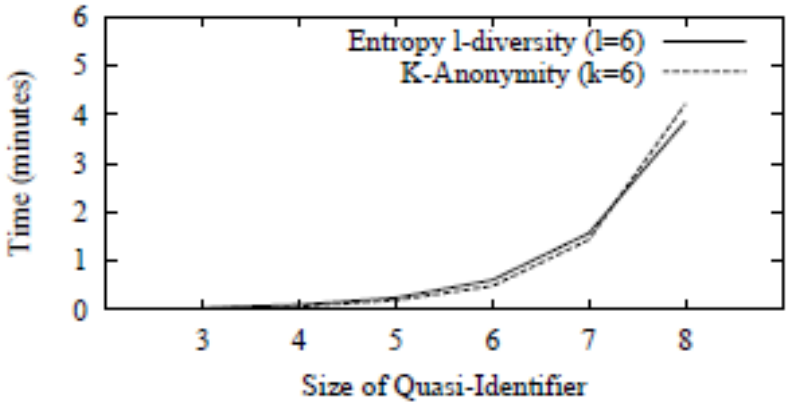
- The algorithm for publishing should find the point on the lattice where the table T* preserves privacy and is useful as possible.
- The usefulness (utility) of table T* is diminished as the data becomes more generalized, so the most utility is at the bottom of the lattice.

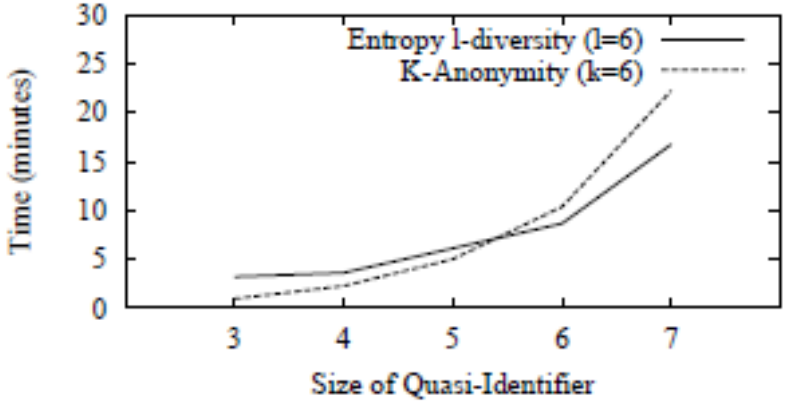
Monotonicity Property

- Monotonicity property is described as a stopping point in the lattice search where the privacy is protected and further generalization does not increase privacy.
- An example is if zip 13065 can be generalized to 1306* and it preserves privacy, generalizing it to 130** also preserves privacy. However, the additional generalization reduces utility.

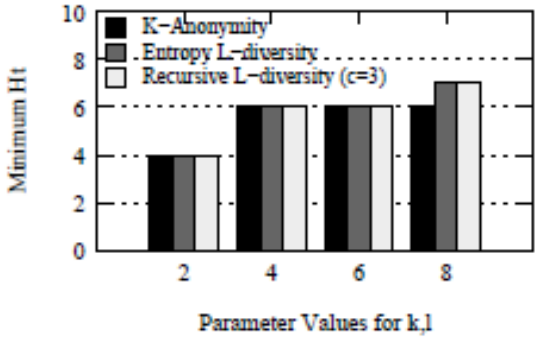
Comparison Testing

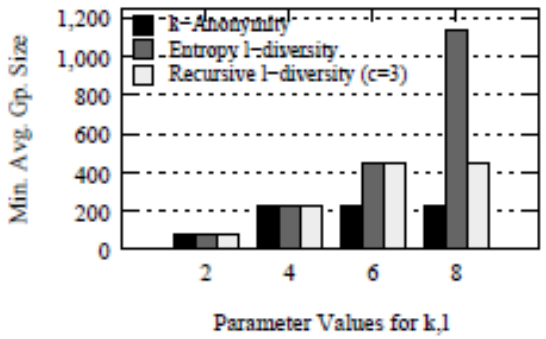
In tests comparing k-anonymization with l-diversity, entropy l-diversity displayed similar if not better run times. As the size of the quasi-identifier grows l-diversity performs better.

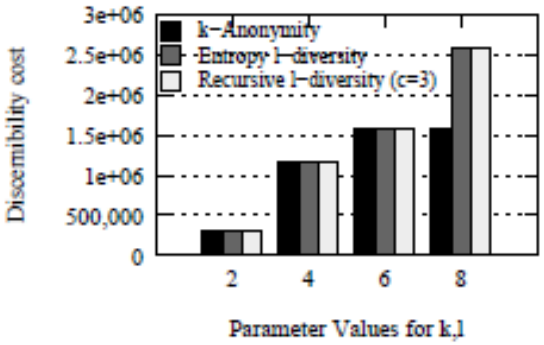




Utility







Using three metrics for utility:

- Generalization height of the anonymized table,
- Minimum average group size of the q*-block
- Discernibility metric is the number of tuples indistinguishable from each other

➢ Smaller values for utility metrics represent higher utility.

Conclusions

The paper presents l-diversity as a means of anonymizing data. They have shown that the algorithms provide a stronger level of privacy than k-anonymity routines.

They have also shown data that supports the claim that the performance and utility differences are minor.

THANK YOU!!