Trường ĐH Bách Khoa Tp.HCM
Khoa Khoa Học và Kỹ Thuật Máy Tính

Privacy Preserving Data Mining: l-diversity

GV: PGS.TS ĐẶNG TRẦN KHÁNH

Chu Xuân Tình -1870583
Nguyễn Đức Huy -1870567

Outline

- Introduction
- Attacks on k-Anonymity
- Bayes Optimal Privacy
- l-Diversity Principle
- Conclusion

Introduction

- 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"

  - 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

Classification of Attributes

- 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

- Each released record should be indistinguishable from at least (k-1) others on its QI attributes
- Alternatively: cardinality of any query result on released data should be at least k
- k-anonymity is (the first) one of many privacy definitions in this line of work
  - l-diversity, t-closeness

  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.

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.

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

Bayes-Optimal Privacy

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

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

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

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

Privacy Principles

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.*

Privacy Principles

Positive Disclosure: Publishing the table T ⋆ 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 ⋆ 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)

The l-Diversity Principle

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:

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

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

The l-Diversity Principle

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*.

The l-Diversity Principle

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.

f (s'|q)=0 (∵ The probability that Umeko has heart disease given her non sensitive attribute as 'Japanese' is 0).
Also, f(s'|q*)=2/4
f(s'|q)/f(s'|q*)=0.

L-Diversity Principle

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

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

L-Diversity Principle

Distinct *l*-diversity – The simplest definition ensures that at least *l* distinct values for the sensitive field in each equivalence class exist.

Entropy *l*-diversity – The most complex definition defines *Entropy* of an equivalent class *E* to be the negation of summation of *s* across the domain of the sensitive attribute of $p(E,s)\log(p(E,s))$ where $p(E,s)$ is the fraction of records in *E* that have the sensitive value *s*. A table has entropy *l*-diversity when for every equivalent class *E*, $Entropy(E) \geq \log(l)$.

Recursive (*c*-*l*)-diversity – A compromise definition that ensures the most common value does not appear too often while less common values are ensured to not appear too infrequently.

L-Diversity Principle

A table is Entropy l- Diverse *if for every* q⋆*-block*

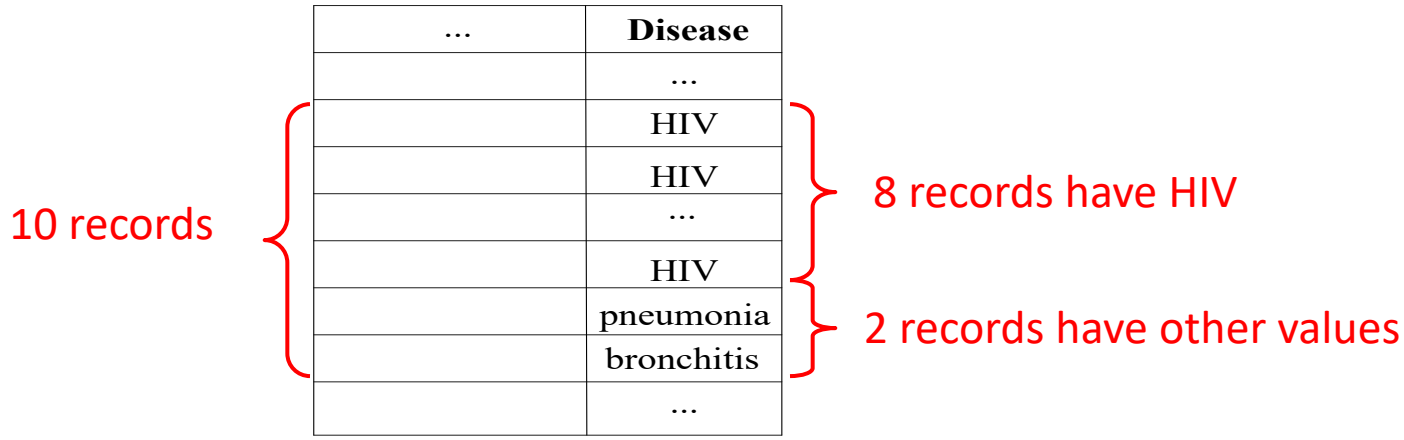
$$-\sum_{s \in S} p_{(q^\star, s)} \log(p_{(q^\star, s')}) \geq \log(\ell)$$

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

Distinct l-Diversity

Each equivalence class has at least l well-represented sensitive values

Doesn't prevent probabilistic inference attacks



Multiple Sensitive Attributes

Definition 4.5 (Multi-Attribute ℓ-Diversity) *Let T be a table with nonsensitive attributes $Q_1, \ldots, Q_{m_1}$ and sensitive attributes $S_1, \ldots, S_{m_2}$. We say that T is ℓ-diverse if for all $i = 1 \ldots m_2$, the table T is ℓ-diverse when S_i is treated as the sole sensitive attribute and $\{Q_1, \ldots, Q_{m_1}, S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_{m_2}\}$ is treated as the quasi-identifier.*

Implementing l-diversity

generalization lattice

assume domain hierarchies exist for all QI attributes

Z2 ={537**}
↑
Z1 ={5371*, 5370*}
↑
Z0 ={53715, 53710, 53706, 53703}

zipcode

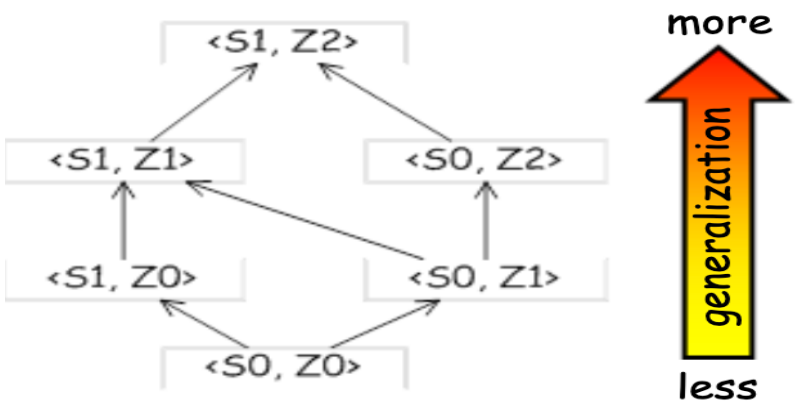
B1 ={*}
↑
B0 ={26/3/1979, 11/3/1980, 16/5/1978}

birthdate

S1 ={Person}
↑
S0 ={Male, Female}

sex

construct the generalization lattice for the entire QI set



Limitations of l-diversity

- l-diversity not prevent attribute disclosure, when multiple records in the table corresponds to one individual.
- *l*-diversity is vulnerable to *skewness attacks* and *similarity attacks*.

Limitations of 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

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.

Limitations of 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!

Limitations of l-diversity

- This leakage of sensitive information occurs because while l-diversity requirement ensures “diversity” of sensitive values in each group, it does not take into account the semantical closeness of these values.
- Summary In short, distributions that have the same level of diversity may provide very different levels of privacy, be- cause there are semantic relationships among the attribute values, because different values have very different levels of sensitivity, and because privacy is also affected by the relationship with the overall distribution.

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.
- l-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. We propose a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table

References

[1] J. Vaidya, C. Clifton, M. Zhu - *Privacy-Preserving Data Mining*. Springer-Verlag, 2006

[2] l-Diversity: Privacy Beyond k-Anonymity.

[3] Internet.

THANK YOU!