Privacy PreservingPublication

Attack and Prevention Models

C21803 Spring 2019

Akash Yadav

Visiting Assistant Professor

Dept. Of CSE, MNNIT Allahabad

Material from the following papers

- Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998

- L-Diversity: Privacy beyond K-Anonymity – Ashwin Machanavajjhala et al., 2006 - (Main Paper for this talk)

Outline

- Defining Privacy
- Need for Privacy
- Source of Problem
- K-anonymity
  - Ways of achieving k-anonymity
    - Generalization
    - Suppression
  - K-minimal Generalizations
- L-diversity
  - K-anonymity attack
  - Primary reasons
  - Model and Notation
  - Bayes Optimal Privacy
  - L-diversity Principle
  - Various Flavours
  - Implementation
  - Experiments

Defining Privacy

- Privacy here means the ***logical security*** of data
- NOT the traditional security of data e.g. access control, theft, hacking etc.
- Here, adversary uses legitimate methods
- Various databases are published e.g. Census data, Hospital records
  - Allows researchers to effectively study the correlation between various attributes

Need for Privacy

- Suppose a hospital has some person-specific patient data which it wants to publish
- It wants to publish such that:
  - Information remains practically useful
  - Identity of an individual cannot be determined
- Adversary might ***infer*** the secret/sensitive data from the published database

Need for Privacy

- The data contains:
  - Attribute values which can uniquely identify an individual { zip-code, nationality, age } or/and {name} or/and {SSN}
  - sensitive information corresponding to individuals { medical condition, salary, location }

	Non-Sensitive Data			*Sensitive Data*	
#	***Zip***	***Age***	***Nationality***	***Name***	***Condition***
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Need for Privacy

Published Data

	Non-Sensitive Data			*Sensitive Data*
#	**Zip**	**Age**	**Nationality**	**Condition**
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Data leak!

#	Name	Zip	Age	Nationality
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Voter List

Source of Problem

- Even if we remove the direct uniquely identifying attributes
  - There are some fields that may still uniquely identify some individual!
  - The attacker can *join* them with other sources and identify individuals

	Non-Sensitive Data			*Sensitive Data*
#	**Zip**	**Age**	**Nationality**	**Condition**
…	…	…	…	…

Quasi-Identifiers

K-anonymity

- Proposed by Sweeney
- Change data in such a way that for each tuple in the resulting table there are atleast (*k-1*) other tuples with the same value for the quasi-identifier – **K-anonymized table**

#	***Zip***	***Age***	***Nationality***	***Condition***
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized

Techniques for anonymization

- Data Swapping
- Randomization
- Generalization
  - Replace the original value by a semantically consistent but *less* specific value
- Suppression
  - Data not released at all
  - Can be Cell-Level or (more commonly) Tuple-Level

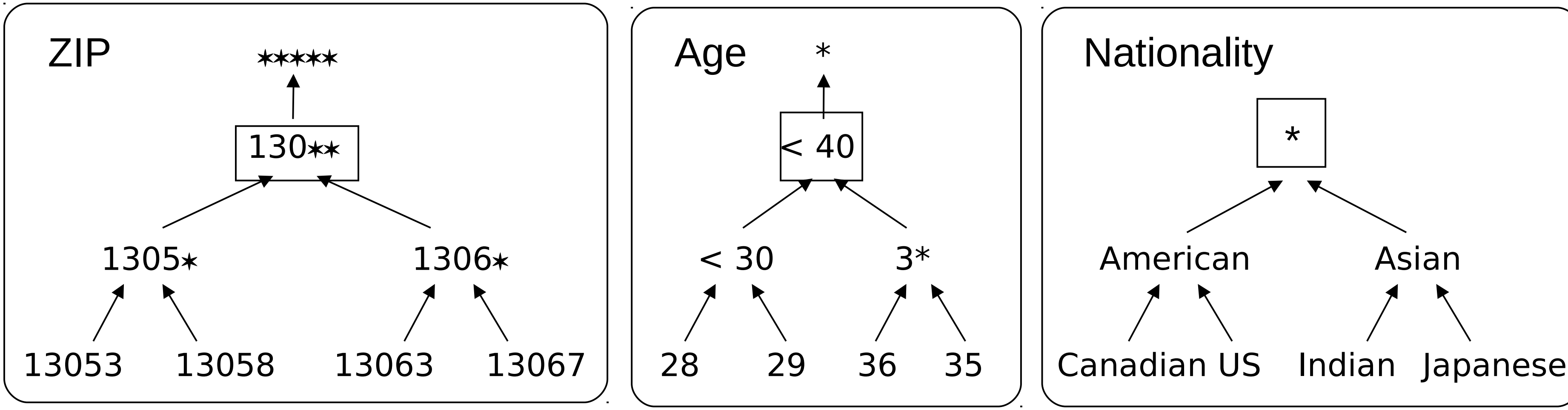
Techniques for anonymization

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Generalization

Suppression (cell-level)

Generalization Hierarchies



- **Generalization Hierarchies**: Data owner defines how values can be generalized
- **Table Generalization**: A table generalization is created by generalizing all values in a column to a specific level of generalization

K-minimal Generalizations

- There are many k-anonymizations – which *one* to pick?
  - Intuition: The one that does not generalize the data more than needed (decrease in utility of the published dataset!)
- **K-minimal generalization**: A k-anonymized table that is not a generalization of another k-anonymized table

#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13053	< 40	*	Viral Infection
3	13067	< 40	*	Heart Disease
4	13067	< 40	*	Cancer

2-minimal Generalizations

#	Zip	Age	Nationality	Condition
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Viral Infection
3	130**	3*	Asian	Heart Disease
4	130**	3*	Asian	Cancer

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Viral Infection
3	130**	< 40	*	Heart Disease
4	130**	< 40	*	Cancer

NOT a 2-minimal Generalization

K-minimal Generalizations

- Now, there are many k-minimal generalizations! – which one is *preferred* then?
- No clear and "correct" answer. It can be
  - The one that creates min. *distortion* to data, where distortion

$$\mathbf{D} = \frac{\sum_{\text{attrib } i} \dfrac{\text{Current level of generalization for attribute } i}{\text{Max level of generalization for attribute } i}}{\text{Number of attributes}}$$

  - The one with min. *supression* i.e. which contains a greater number of tuples *and so on*

Complexity & Algorithms

- If we allow for generalization to a different level for each value of an attribute, the search space is exponential
- More often than not, the problem is NP-Hard!
- Many algorithms have been proposed
  - Incognito
  - Multi-dimensional algorithms (Mondrian)

K-Anonymity Drawbacks

- K-anonymity alone *does not* provide full privacy!
- Suppose attacker knows the non-sensitive attributes of

	Zip	***Age***	***National***
Bob	13053	31	American
Umeko	13068	21	Japanese

- And the fact that Japanese have very low incidence of heart disease

K-Anonymity Attack

Original Data

	Non-Sensitive Data			*Sensitive Data*
#	**ZIP**	**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

4-anonymized Table

Umeko Matches here

Bob Matches here

	Non-Sensitive Data			*Sensitive Data*
#	**ZIP**	**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

4-anonymized Table

	Non-Sensitive Data			Sensitive Data
#	***ZIP***	***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	[illegible]	[illegible]	[illegible]	*Viral Infection*
9	[illegible]	[illegible]	[illegible]	Cancer
10	130**	3*	*	Cancer

Umeko Matches here

Bob Matches here

Bob has Cancer!

4-anonymized Table

	Non-Sensitive Data			*Sensitive Data*
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130[illegible]	[illegible]	[illegible]	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	[illegible]	[illegible]	[illegible]	*Viral Infection*
9	130[illegible]	3[illegible]	[illegible]	Cancer
10	130**	3*	*	Cancer

Umeko Matches here

Umeko has Viral Infection!

Bob Matches here

Bob has Cancer!

K-Anonymity Drawbacks

- Basic Reasons for leak –
  - Sensitive attributes lack *diversity* in values
    - Homogeneity Attack
  - Attacker has additional *background knowledge*
    - Background knowledge Attack

- Hence a new solution has been proposed *in-addition* to k-anonymity – *l-diversity*

L-diversity

- Proposed by Ashwin M. et al. SIGMOD 2006
- **Model and notation**:

$$T = \{t_1, t_2, \ldots, t_n\} \qquad A_1, A_2, \ldots, A_m$$

$\Omega =$ population from which T has been taken

$t[C] = (t[C_1, C_2, \ldots, C_p])$ where C is a set

$S =$ set of Sensitive attrib ; $QI =$ set of Quasi-identifiers

$T^* =$ Anonymized table

Model and Notation

- As a sanity check to understand all the notation ☺, here is a simple definition of k-anonymity

Definition **(k-Anonymity)** *A table* T *satisfies* k*-anonymity if for every tuple* $t \in T$ *there exist* $k-1$ *other tuples* $t_{i_1}, t_{i_2}, \ldots, t_{i_{k-1}} \in T$ *such that* $t[\mathcal{C}] = t_{i_1}[\mathcal{C}] = t_{i_2}[\mathcal{C}] = \cdots = t_{i_{k-1}}[\mathcal{C}]$ *for all* $\mathcal{C} \in \mathcal{QI}$.

- Consider only generalization techniques for k-anonymity

Model and Notation

- Adversary’s Background Knowledge
  - Has access to published table *T** and knows that it is a generalization of some base table *T*
  - May also know that some individuals are present in the table. E.g. Alice may know Bob has gone to the hospital -> his records will be present
  - May also have partial knowledge about the distribution of sensitive and non-sensitive attribs. in the population

Bayes Optimal Privacy

- Ideal Notion of privacy
- Models background knowledge as probability distribution over attributes
- Uses Bayesian Inference techniques
- Assume, T is a simple random sample and only a single sensitive attribute S and a condensed quasi-identifier attribute Q
- Assume worst case, adversary (Alice) knows the complete joint distribution f of Q and S

Bayes Optimal Privacy

- Alice has a *prior belief* of (say) Bob's sensitive attribute (given his *Q* attributes) i.e.

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

- After *T** Alice's belief changes to its *posterior value* i.e.

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

- Given *f* and *T** we can calculate the posterior

Bayes Optimal Privacy

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

The proof is involved. See extended paper for proof.

$n_{(q^\star,s')}$ is the number of tuples in *T** with the *t*[Q]* = *q** and *t*[S]* = *s'*

Bayes Optimal Privacy

Definition (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; i.e., given a $\delta > 0$, there is a positive disclosure if $\beta_{(q,s,T^\star)} > 1 - \delta$ and there exists $t \in T$ such that $t[Q] = q$ and $t[S] = s$.*

Definition (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); i.e., given an $\epsilon > 0$, there is a negative disclosure if $\beta_{(q,s,T^\star)} < \epsilon$ and there exists a $t \in T$ such that $t[Q] = q$ but $t[S] \neq s$.*

Bayes Optimal Privacy

- Note not all p.d.s and n.d.s are bad
  - If Alice already knew Bob has Cancer, there is nothing much one can do!
- Hence, intuitively, there should not be a large difference in the prior and posterior
- Different privacy breach metrics
- Note that diversity and background knowledge are both captured in any definition!

Bayes Optimal Privacy

- Limitations in practice
  - Data publisher unlikely to know *f*
  - Publisher does not know how much the adversary actually knows
    - He may have instance level knowledge
    - No way to model non-probabilistic knowledge
  - Multiple adversaries having different levels of knowledge
- Hence a *practical* definition is needed

L-diversity principle

- Consider p.d.s : Alice wants to determine Bob’s sensitive attrib. with high probability
- Using posterior, can happen only when

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

- Which in turn can occur due to both lack of diversity and/or background knowledge

L-diversity principle

- Lack of diversity manifests as

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

- This can guarded against by requiring "many" sensitive values are "well-represented" in a *q** block (a generalization block)
- Background Knowledge

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

L-diversity principle

- Note that Alice has to *eliminate* other sensitive values to get a p.d.
- But if *l* values are "well-represented", Alice intuitively needs at least *l-1* damaging pieces of information!
- Hence, we get a practical principle:

Principle **(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.*

3-diverse Table

	Non-Sensitive Data			*Sensitive Data*
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	1305*	<= 40	*	Heart Disease
2	1305*	<= 40	*	Viral Infection
3	1305*	<= 40	*	Cancer
4	1305*	<= 40	*	Cancer
5	*1485**	*>= 40*	*	*Cancer*
6	*1485**	*>= 40*	*	*Heart Disease*
7	*1485**	*>= 40*	*	*Viral Infection*
8	*1485**	*>= 40*	*	*Viral Infection*
9	1306*	<= 40	*	Heart Disease
10	1306*	<= 40	*	Viral Infection
11	1306*	<= 40	*	Cancer

Some L-diversity Instantiations

- Entropy L-Diversity

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

$$where\ p_{(q^\star,s)} = \frac{n_{(q^\star,s)}}{\sum_{s'\in S} n_{(q^\star,s')}}$$

Some L-diversity Instantiations

- Need the entropy of original table at least *log(l)*
  - Too restrictive
  - One value of sensitive attr. may be very common
- Recursive *(c, l)*-Diversity
  - None of the sensitive values should occur *too* frequently.
  - Let r_i be the i^{th} most frequent sensitive value
- Given const *c*, satisfies *(c, l)* diversity if

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

Some L-diversity Instantiations

- Positive Disclosure-Recursive *(c, l)*-Diversity

Let Y denote the set of sensitive values for which positive disclosure is allowed. In a given $q^\star$-block, let the most frequent sensitive value not in Y be the y^{th} most frequent sensitive value. Let r_i denote the frequency of the i^{th} most frequent sensitive value in the $q^\star$-block. Such a $q^\star$-block satisfies pd-recursive (c, ℓ)-diversity *if one of the following hold:*

- $y \leq \ell - 1$ *and* $r_y < c \sum_{j=\ell}^{m} r_j$
- $y > \ell - 1$ *and* $r_y < c \sum_{j=\ell-1}^{y-1} r_j + c \sum_{j=y+1}^{m} r_j$

Some L-diversity Instantiations

- Negative/Positive Disclosure-Recursive (c_1, c_2, l)-Diversity
  - Consider n.d.s also
  - Let *W* be set of sensitive values for which n.d.s are not allowed
  - Requirement
    - Pd-recursive (c_1, l)
    - Every *s* in *W* occurs at least c_2 percent of tuples in every block

Multiple Sensitive Attributes

- Recall we assumed a *single* sensitive attribute *S*
- What if there are 2 sensitive attrib *S* and *V*?
  - It may individually be l-diverse
  - But, as a whole, it may violate
    - *V* may not be well-represented for each value of *S*
  - Solution
    - Include *S* in the quasi-identifier set when checking for diversity in *V*
    - And vice versa! – Easy to generalize

Implementation

- Most k-anonymization algos search the generalization space
  - Recall, in general it is *NP*-Hard
  - Can be made more efficient if the *Monotonicity* condition holds
    - If *T** preserves privacy, then so does every generalization of it
  - If l-diversity also possesses this property
    - We can re-use previous algos directly
    - Whenever we check for k-anon., check for l-diversity instead
  - Fortunately! All flavours except the Bayes Optimal Privacy is monotonic

Experiments

- Used Incognito (a popular generalization algorithm)

Adults Database Description

Adults

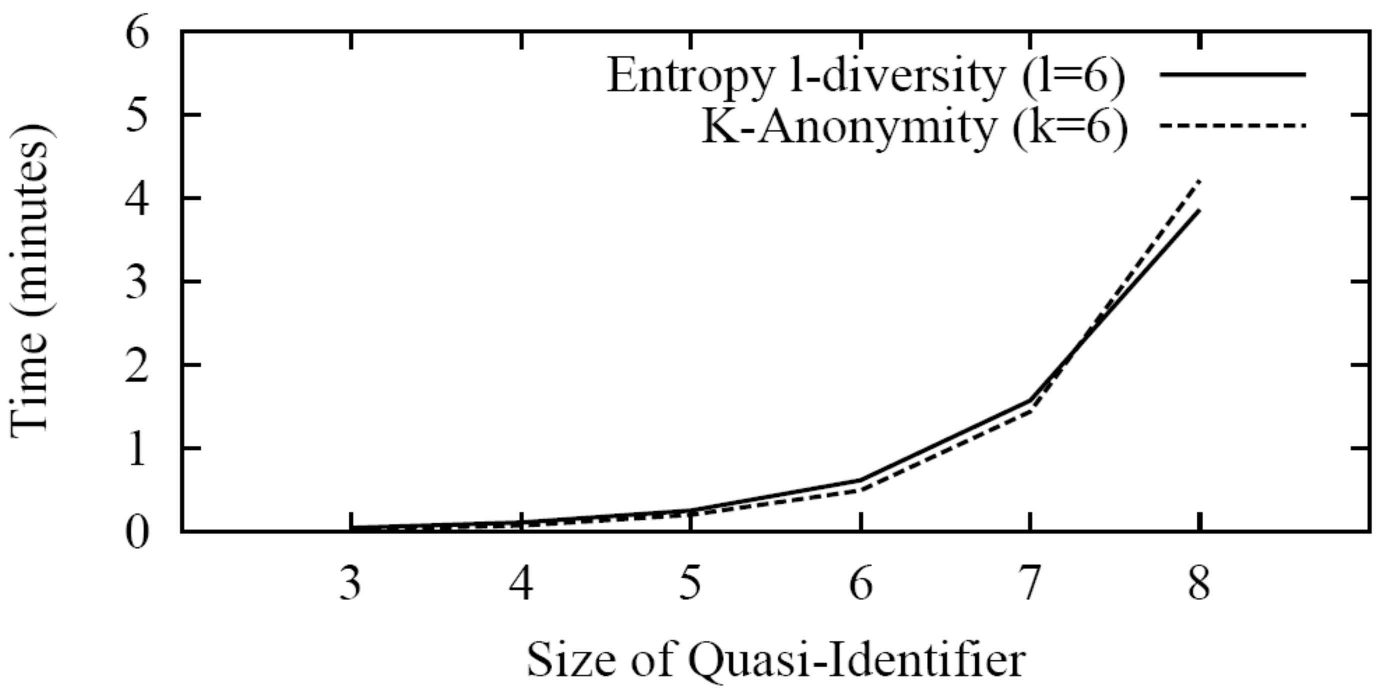
	Attribute	Domain size	Generalizations type	Ht.
1	Age	74	ranges-5,10,20	4
2	Gender	2	Suppression	1
3	Race	5	Suppression	1
4	Marital Status	7	Taxonomy tree	2
5	Education	16	Taxonomy tree	3
6	Native Country	41	Taxonomy tree	2
7	Work Class	7	Taxonomy tree	2
8	Salary class	2	*Sensitive att.*	
9	Occupation	41	*Sensitive att.*	

Experiments

- Homogeneity Attack
  - Treat first 5 attributes as quasi-identifier, Occupation as sensitive attirb.
    - 12 minimal 6-anon. tables generated, one was vulnerable
    - If Salary is sensitive attrib, out of 9 minimal 6-anon., 8 were prone to attack
  - So, homogeneity attack prone k-anonymized datasets are routinely produced

Experiments

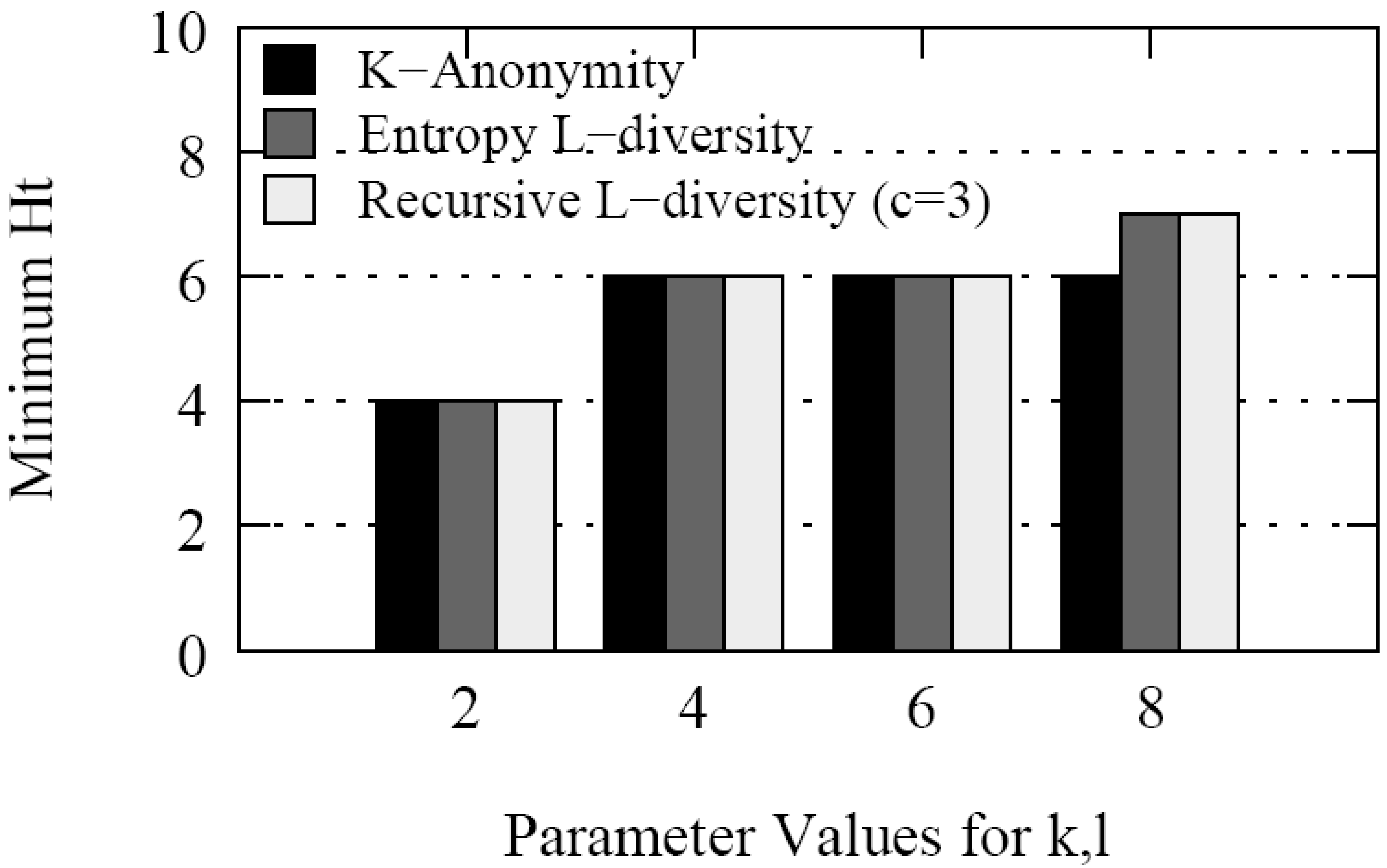
- Performance
  - Does l-diversity incur heavy overhead?
    - Comparing time to return 6-diverse Vs 6-anon. tables



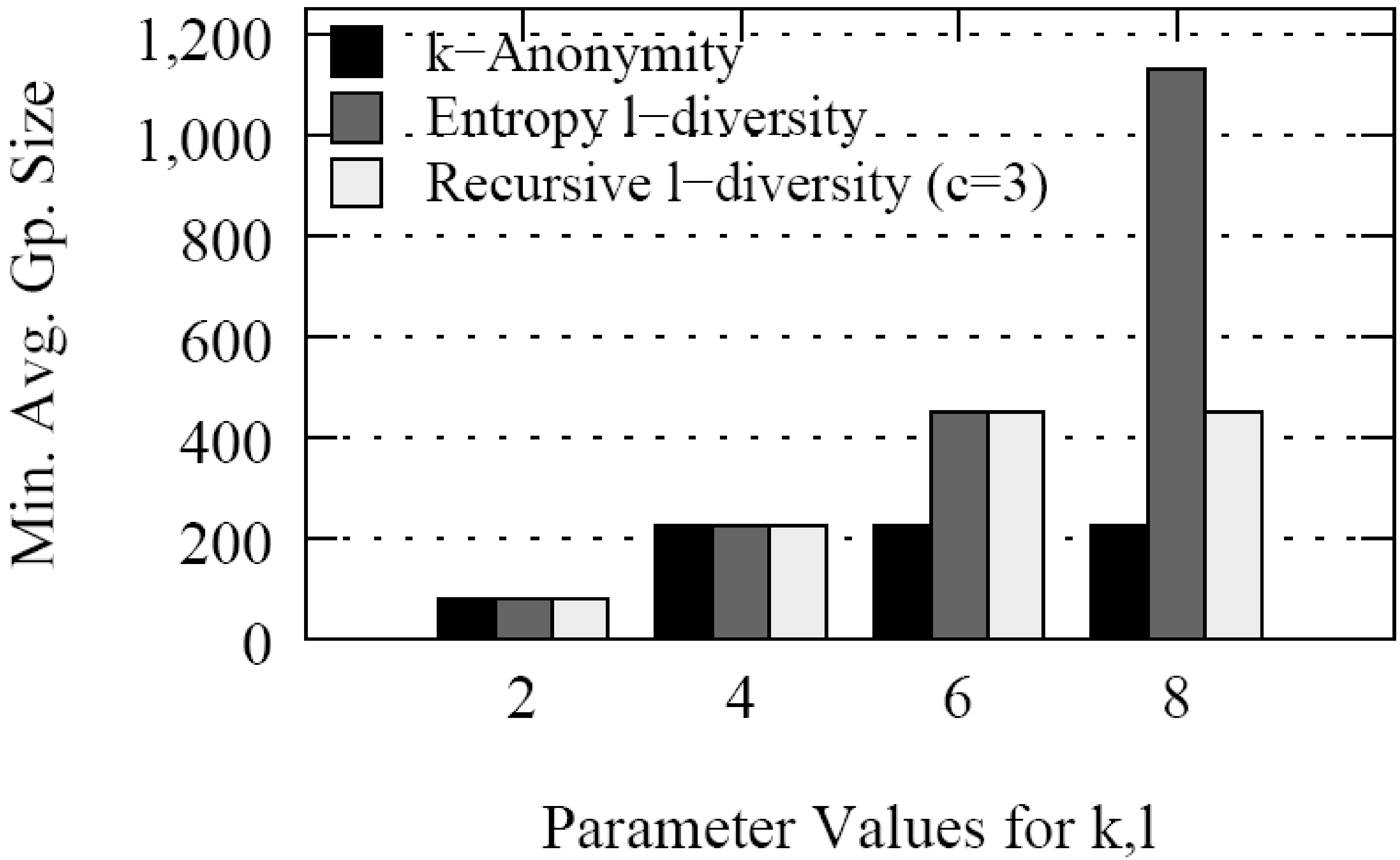
Experiments

- Utility
  - Intuitively: “usefulness” of the l-diverse and k-anonymized tables
    - No clear metric
    - Used 3 different metrics
      - No. of generalization steps that were performed
      - Average size of *q**-blocks generated
      - Discernibility Metric - Measures the no. of tuples indistinguishable from each other
  - Used *k, l* = 2, 4, 6, 8

Experiments



Experiments



Thank You!

- Any Questions?