Privacy Skyline:

Privacy with Multidimensional Adversarial Knowledge

Bee-Chung Chen, Kristen LeFevre
University of Wisconsin – Madison

Raghu Ramakrishnan
Yahoo! Research

Example: Medical Record Dataset

- A data owner wants to release data for medical research
- An adversary wants to discover individuals' sensitive info

Name	Age	Gender	Zipcode	Disease
Ann	20	F	12345	AIDS
Bob	24	M	12342	Flu
Cary	23	F	12344	Flu
Dick	27	M	12343	AIDS
Ed	35	M	12412	Flu
Frank	34	M	12433	Cancer
Gary	31	M	12453	Cancer
Tom	38	M	12455	AIDS

What If the Adversary Knows …

	Age	Gender	Zipcode	Group
(Ann)	2*	Any	1234*	1
(Bob)				
(Cary)				
(Dick)				
(Ed)	3*	M	123**	2
(Frank)				
(Gary)				
(Tom)				

Group	Disease
1	AIDS Flu Flu AIDS
2	Flu Cancer Cancer AIDS

- Without any additional knowledge, Pr(Tom has AIDS) = ¼
- What if the adversary knows "Tom does not have Cancer and Ed has Flu"

Pr(Tom has AIDS | above data and above knowledge) = **1**

Privacy with Adversarial Knowledge

- Bayesian privacy definition: A released dataset $\mathbf{D}^*$ is **safe** if, for any person t and any sensitive value s,

$$\Pr(\ t \text{ has } s \mid \mathbf{D}^*, \text{Adversarial Knowledge}\ ) < c$$

  - This probability is the adversary's confidence that person t has sensitive value s, after he sees the released dataset
  - Equivalent definition: $\mathbf{D}^*$ is safe if

$$\underbrace{\max_{t,s} \Pr(t \text{ has } s \mid \mathbf{D}^*, \text{Adversarial Knolwedge})}_{\text{Maximum breach probability}} < c$$

  - Prior work following this intuition: [Machanavajjhala et al., 2006; Martin et al., 2007; Xiao and Tao, 2006]

Questions to be Addressed

- Bayesian privacy criterion:

  max Pr(t has s | $\mathbf{D}^*$, Adversarial Knowledge) < c

- **How to describe various kinds of adversarial knowledge**
  - We provide intuitive knowledge expressions that cover three kinds of common adversarial knowledge
- **How to analyze data safety in the presence of various kinds of possible adversarial knowledge**
  - We propose a skyline tool for what-if analysis in the "knowledge space"
- **How to efficiently generate a safe dataset to release**
  - We develop algorithms (based on a "congregation" property) orders of magnitude faster than the best known dynamic programming technique [Martin et al., 2007]

Outline

- **Theoretical framework (possible-world semantics)**
  - **How the privacy breach is defined**
- Three-dimensional knowledge expression
- Privacy Skyline
- Efficient and scalable algorithms
- Experimental results
- Conclusion and future work

Theoretical Framework

Original dataset **D**

Name	Age	Gender	Zipcode	Disease
Ann	20	F	12345	AIDS
Bob	24	M	12342	Flu
Cary	23	F	12344	Flu
Dick	27	M	12343	AIDS
Ed	35	M	12412	Flu
Frank	34	M	12433	Cancer
Gary	31	M	12453	Cancer
Tom	38	M	12455	AIDS

- Assume each person has only one sensitive value (in the talk)
- Sensitive attribute can be set-valued (in the paper)

Release candidate **D***

	Age	Gender	Zipcode	Group
(Ann)	20	F	12345	1
(Bob)	24	M	12342	
(Cary)	23	F	12344	
(Dick)	27	M	12343	
(Ed)	35	M	12412	2
(Frank)	34	M	12433	
(Gary)	31	M	12453	
(Tom)	38	M	12455	

Group	Disease
1	AIDS Flu Flu AIDS
2	Flu Cancer Cancer AIDS

- Each group is called a QI-group
- This abstraction includes
  - Generalization-based methods
  - Bucketization

Theoretical Framework

Reconstruction

A reconstruction of $\mathbf{D}^*$ is intuitively a possible original dataset (possible world) that would generate $\mathbf{D}^*$ by using the grouping mechanism

Release candidate $\mathbf{D}^*$

	Age	Gender	Zipcode	Group
(Ann)	20	F	12345	1
(Bob)	24	M	12342	1
(Cary)	23	F	12344	1
(Dick)	27	M	12343	1
(Ed)	35	M	12412	2
(Frank)	34	M	12433	2
(Gary)	31	M	12453	2
(Tom)	38	M	12455	2

Group	Disease
1	AIDS Flu Flu AIDS
2	Flu Cancer Cancer AIDS

Reconstructions of Group 2

Ed	…	Flu
Frank	…	Cancer
Gary	…	Cancer
Tom	…	AIDS

⋮

Ed	…	AIDS
Fra[illegible]	…	[illegible]
Gra[illegible]	…	[illegible]
Tom	…	Flu

Fix Permute

Assumption: Without any additional knowledge, every reconstruction is equally likely

Probability Definition

- Knowledge expression K: Logic sentence [Martin et al., 2007]

  E.g., $K = (\text{Tom}[S] \neq \text{Cancer}) \wedge (\text{Ed}[S] = \text{Flu})$

  $$\Pr(\ \text{Tom}[S] = \text{AIDS} \mid K, \mathbf{D}^*\ )$$

  $$\equiv \frac{\#\text{ of reconstructions of } \mathbf{D}^* \text{ that satisfy } K \wedge (\text{Tom}[S] = \text{AIDS})}{\#\text{ of reconstructions of } \mathbf{D}^* \text{ that satisfy } K}$$

- Worst-case disclosure
  - Knowledge expressions may also include variables

    E.g., $K = (\text{Tom}[S] \neq \boldsymbol{x}) \wedge (\boldsymbol{u}[S] \neq \boldsymbol{y}) \wedge (\boldsymbol{v}[S] = \boldsymbol{s} \rightarrow \text{Tom}[S] = \boldsymbol{s})$
  - Maximum breach probability

    $$\max \Pr(\ \boldsymbol{t}[S] = \boldsymbol{s} \mid \mathbf{D}^*, K\ )$$

    The maximization is over variables t, u, v, s, x, y, by substituting them with constants in the dataset

What Kinds of Expressions

- Privacy criterion: Release candidate $\mathbf{D}^*$ is safe if

$$\max \Pr(\ t[S] = s \mid \mathbf{D}^*, K\ ) < c$$

- Prior work by Martin et al., 2007
  - K is a conjunction of m implications

    E.g., $K = (u_1[S] = x_1 \rightarrow v_1[S] = y_1) \wedge \ldots \wedge (u_m[S] = x_m \rightarrow v_m[S] = y_m)$
  - Not intuitive: What is the practical meaning of m implications?
  - Some limitations: Some simple knowledge cannot be expressed
- Complexity for general logic sentences
  - Computing breach probability is NP-hard
- Goal: Identify classes of expressions that are
  - Useful (intuitive & cover common adversarial knowledge)
  - Computationally feasible

Outline

- Theoretical framework
- **Three-dimensional knowledge expression**
  - **Tradeoff between expressiveness and feasibility**
- Privacy Skyline
- Efficient and scalable algorithms
- Experimental results
- Conclusion and future work

Kinds of Adversarial Knowledge

	Age	Gender	Zipcode	Group
(Ann)	20	F	12345	1
(Bob)	24	M	12342	
(Cary)	23	F	12344	
(Dick)	27	M	12343	
(Ed)	35	M	12412	2
(Frank)	34	M	12433	
(Gary)	31	M	12453	
(Tom)	38	M	12455	

Group	Disease
1	AIDS Flu Flu AIDS
2	Flu Cancer Cancer AIDS

Assume a person has only one record in the dataset in this talk
(Multiple sensitive values per person is in the paper)

- Adversary's target: Whether Tom has AIDS
- **Knowledge about the target:** Tom does not have Cancer
- **Knowledge about other people:** Ed has Flu
- **Knowledge about relationships:** Ann has the same sensitive value as Tom

3D Knowledge Expression

- Adversary's target: Whether person t has sensitive value s
- Adversary's knowledge $\mathscr{L}_{t,s}(\ell,k,m)$:
  - **Knowledge about the target: *ℓ* sensitive values** that t does not have

$$t[S] \neq x_1 \wedge \ldots \wedge t[S] \neq x_\ell$$

  - **Knowledge about others:** The sensitive values of ***k* other people**

$$u_1[S] = y_1 \wedge \ldots \wedge u_k[S] = y_k$$

  - **Knowledge about relationships:** A group of ***m* people** who have the same sensitive value as t

$$(v_1[S] = s \rightarrow t[S] = s) \wedge \ldots \wedge (v_m[S] = s \rightarrow t[S] = s)$$

- Worst-case guarantee: max $\Pr(t[S] = s \mid \mathbf{D}^*, \mathscr{L}_{t,s}(\ell,k,m)) < c$
  - No matter what those ***ℓ* sensitive values**, what those ***k* people** and what those ***m* people** are, the adversary should <u>not</u> be able to predict any person t to have any sensitive value s with confidence $\geq c$

Outline

- Theoretical framework
- Three-dimensional knowledge expression
- **Privacy Skyline**
  - **Skyline privacy criterion**
  - **Skyline exploratory tool**
- Efficient and scalable algorithms
- Experimental results
- Conclusion and future work

Basic 3D Privacy Criterion

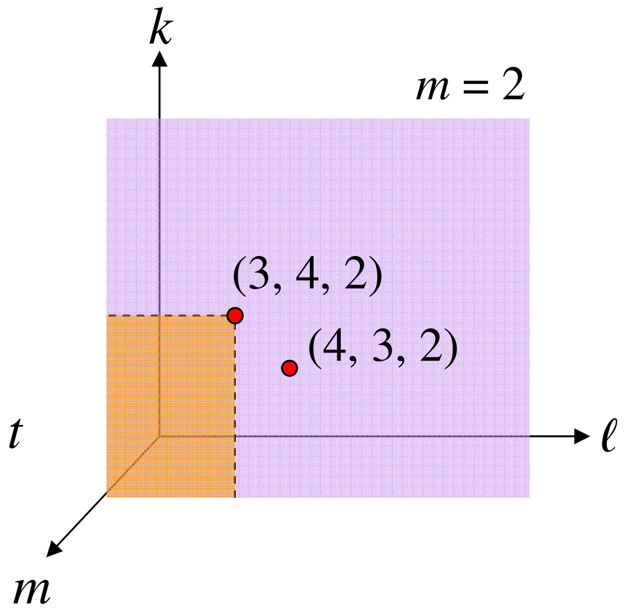
- Given knowledge threshold (ℓ, k, m) and confidence threshold c, release candidate $\mathbf{D}^*$ is **safe** if

$$\max \Pr(\ t[S] = s \mid \mathbf{D}^*, \mathscr{L}_{t,s}(\ell,k,m)\ ) < c$$

Example: (ℓ, k, m) = (3, 4, 2) and c = 0.5

A release candidate is **safe** if no adversary with the following knowledge can predict any person t to have any sensitive value s with confidence ≥ 0.5

- Any 3 sensitive values that t does not have
- The sensitive values of any 4 people
- Any 2 people having the same sensitive value as t



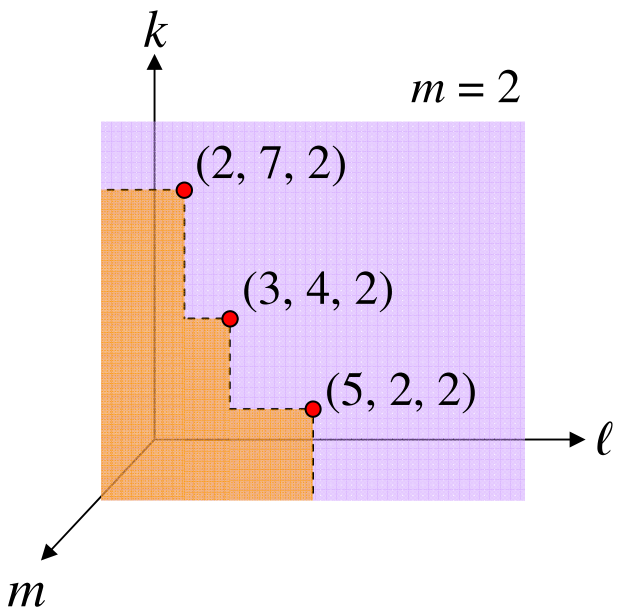
***k*-anonymity and *ℓ*-diversity are two special cases of this criterion**

Skyline Privacy Criterion

- Given a set of skyline points

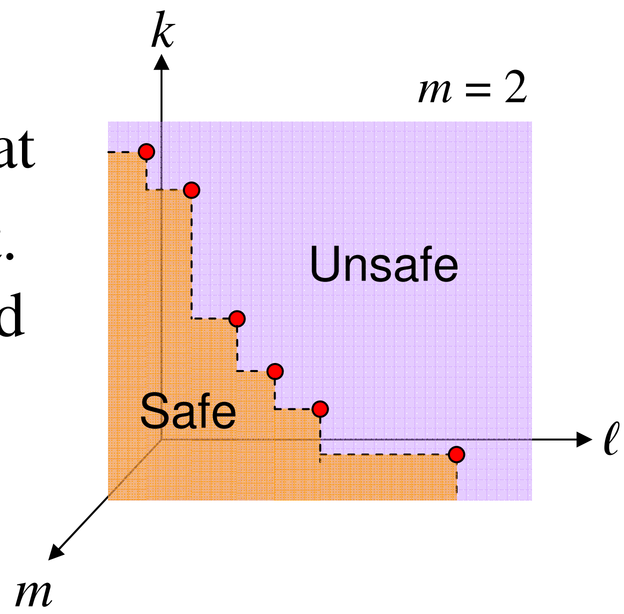
$$(\ell_1, k_1, m_1, c_1), \ldots, (\ell_r, k_r, m_r, c_r),$$

release candidate $\mathbf{D}^*$ is **safe** if it is safe with respect to every point



Skyline Exploratory Tool

- In the skyline privacy criterion
  - The data owner specifies a set of skyline points
  - The system checks whether a release candidate is safe
- Skyline exploratory tool
  - Given a release candidate
  - Find the set of skyline points such that
    - The release candidate is **safe** w.r.t. any point **beneath** the skyline, and
    - The release candidate is **unsafe** w.r.t. any point **above** the skyline



Outline

- Theoretical framework
- Three-dimensional knowledge expression
- Privacy Skyline
- **Efficient and scalable algorithms**
  - **SkylineCheck (in this talk)**
    - **Check whether a given release candidate is safe w.r.t. a skyline**
  - SkylineAnonymize (in the paper)
    - Generate a safe release candidate that maximizes a utility function
  - SkylineFind (in the technical report)
    - Find the skyline of a given release candidate
- Experimental results
- Conclusion and future work

Check Safety for a Single Point

- Given (ℓ, k, m, c), check

$$\max \Pr(\ t[S] = s \mid \mathbf{D}^*, \mathcal{L}_{t,s}(\ell,k,m)\ ) < c$$

  - $\mathcal{L}_{t,s}(\ell,k,m) = K_t(\ell) \wedge K_u(k) \wedge K_{v,t}(m)$
    - $K_t(\ell) = t[S] \neq x_1 \wedge \ldots \wedge t[S] \neq x_\ell$
    - $K_u(k) = u_1[S] = y_1 \wedge \ldots \wedge u_k[S] = y_k$
    - $K_{v,t}(m) = (v_1[S] = s \rightarrow t[S] = s) \wedge \ldots \wedge (v_m[S] = s \rightarrow t[S] = s)$
  - Variables:
    - People: $t, u_1, \ldots, u_k, v_1, \ldots, v_m$
    - Sensitive values: $x_1, \ldots, x_\ell, y_1, \ldots, y_k$
- Technical challenge:
  - How to find the variable assignment that maximizes the breach probability

Check Safety for a Single Point

- max Pr($t[S] = s \mid \mathbf{D}^*, \mathscr{L}_{t,s}(\ell,k,m)$)
  - Variables:
    - People: $t, u_1, \ldots, u_k, v_1, \ldots, v_m$
    - Sensitive values: $x_1, \ldots, x_\ell, y_1, \ldots, y_k$
- In principle, we need to
  - Consider **all possible ways** of assigning **person variables** into QI-groups
  - For each assignment of person variables, find the assignment of **sensitive-value variables** that maximizes the breach probability
    - Has a closed-form solution

Release candidate $\mathbf{D}^*$

Age	Gender	Zipcode	Group
20	F	12345	1
24	M	12342	
23	F	12344	
27	M	12343	
35	M	12412	2
34	M	12433	
31	M	12453	
38	M	12455	
20	F	12345	3
24	M	12342	
23	F	12344	
27	M	12343	
35	M	12412	4
34	M	12433	
31	M	12453	
38	M	12455	

Group	Disease
1	AIDS
	Flu
	Flu
	AIDS
2	Flu
	Cancer
	Cancer
	AIDS
3	AIDS
	Flu
	Flu
	AIDS
4	Flu
	Cancer
	Cancer
	AIDS

Example assignment of person variables:

Group 1: t, u_1
Group 2: u_2, v_1, v_2
Group 3: u_3, u_4
Group 4: v_3, v_4

"Congregation" Property

- max Pr($t[S] = s \mid \mathbf{D}^*, \mathcal{L}_{t,s}(\ell,k,m)$)
  - Variables:
    - People: $t, u_1, \ldots, u_k, v_1, \ldots, v_m$
    - Sensitive values: $x_1, \ldots, x_\ell, y_1, \ldots, y_k$
- When the breach probability is maximized,
  - All $u_1, \ldots, u_k$ would congregate in one QI-group
  - All $v_1, \ldots, v_m$ would congregate in one QI-group
  - t would be in one of the above two

Release candidate $\mathbf{D}^*$

Age	Gender	Zipcode	Group
20	F	12345	1
24	M	12342	1
23	F	12344	1
27	M	12343	1
35	M	12412	2
34	M	12433	2
31	M	12453	2
38	M	12455	2
20	F	12345	3
24	M	12342	3
23	F	12344	3
27	M	12343	3
35	M	12412	4
34	M	12433	4
31	M	12453	4
38	M	12455	4

Group	Disease
1	AIDS
1	Flu
1	Flu
1	AIDS
2	Flu
2	Cancer
2	Cancer
2	AIDS
3	AIDS
3	Flu
3	Flu
3	AIDS
4	Flu
4	Cancer
4	Cancer
4	AIDS

Example assignment of person variables:

Group 1:
Group 2: $t, u_1, \ldots, u_k$
Group 3:
Group 4: $v_1, \ldots, v_m$

Five Sufficient Statistics

- Three possible cases at the maximum
  - Case 1:
    - All person variables are in one QI-group (A)

      $\max \Pr(\ldots) = 1 / [\ (\mathbf{min}_A\ \boldsymbol{CF_1(A)}) + 1]$
  - Case 2:
    - t and $u_1, \ldots, u_k$ are in one QI-group (B)
    - $v_1, \ldots, v_m$ are in one QI-group (C)

      $\max \Pr(\ldots) = 1 / [\ (\mathbf{min}_B\ \boldsymbol{CF_2(B)})\cdot(\mathbf{min}_C\ \boldsymbol{CF_3(C)}) + 1]$
  - Case 3:
    - t and $v_1, \ldots, v_m$ are in one QI-group (D)
    - $u_1, \ldots, u_k$ are in one QI-group (E)

      $\max \Pr(\ldots) = 1 / [\ (\mathbf{min}_D\ \boldsymbol{CF_4(D)})\cdot(\mathbf{min}_E\ \boldsymbol{CF_5(E)}) + 1]$

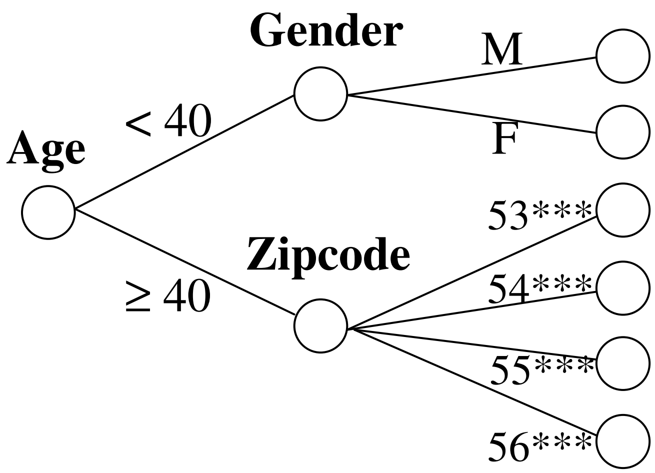
(For a fixed QI-group, $CF_1, \ldots, CF_5$ are closed-form formulas)

SkylineCheck Algorithm

- Keep 5 sufficient statistics (5 floating-point variables) for each skyline point
- Single-scan algorithm
  - Scan the dataset once
  - During the scan, update the 5 sufficient statistics for each skyline point
  - Compute the maximum breach probability based on these statistics

SkylineAnonymize Algorithm

- Goal: Generate a safe release candidate that maximizes a utility function
- Partition records into QI-groups by a tree structure
  - Adaptation of the Mondrian algorithm by LeFevre et al.
  - **The congregation property makes the adaptation easy**



Age	Gender	Zipcode	Group
			1
			2
			3
			4
			5
			6

Group	Disease
1	
2	
3	
3	
5	
6	

Outline

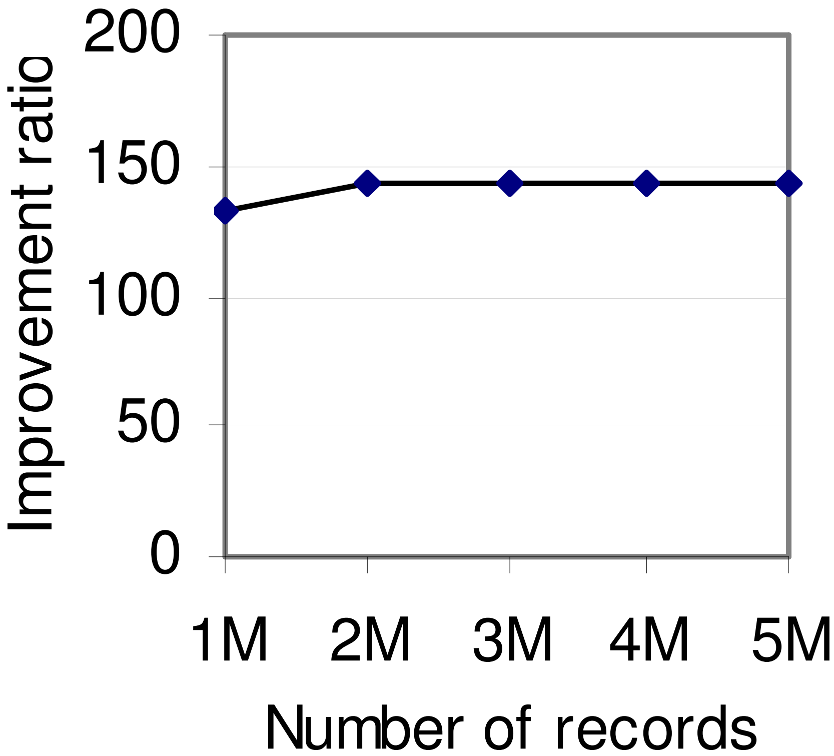
- Theoretical framework
- Three-dimensional knowledge expression
- Privacy Skyline
- Efficient and scalable algorithms
- **Experimental results**
- Conclusion and future work

Experimental Results

- Our SkylineCheck algorithm (based on the congregation property) is orders of magnitude faster than the best-known dynamic-programming technique [Martin et al., 2007]
- Our SkylineAnonymize algorithm scales nicely to datasets substantially larger than main memory
- A case study shows usefulness of the skyline exploratory tool

Efficiency of SkylineCheck

(ℓ=10, k =10, m =10)

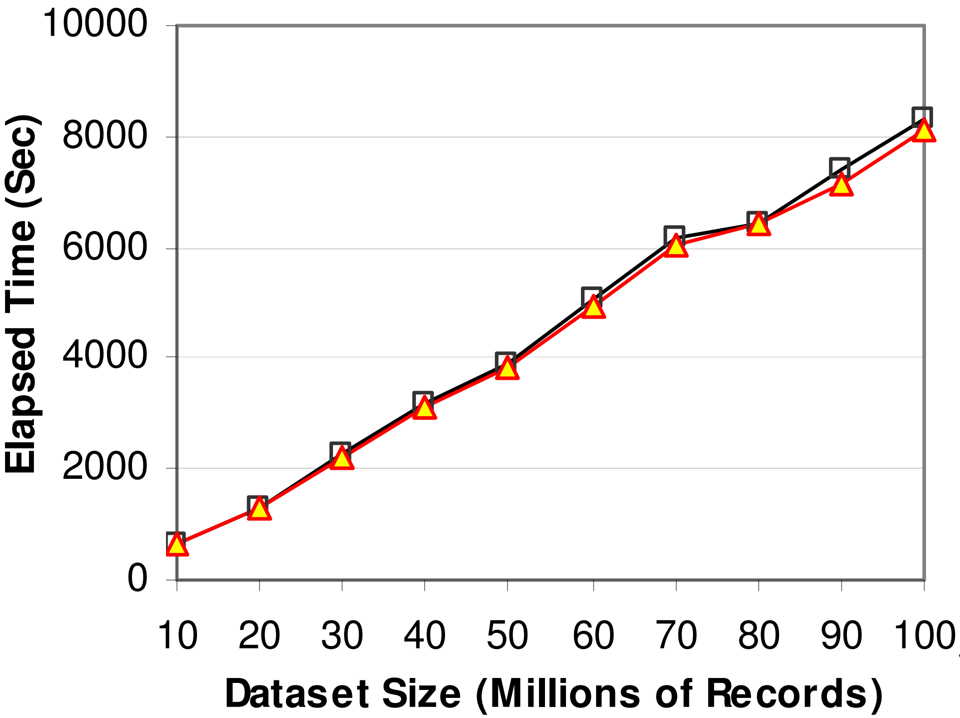


Improvement ratio =

$$\frac{\text{Execution time of DP}}{\text{Execution time of ours}}$$

Scalability of SkylineAnonymize

Main memory size: **512 MB**
Record size: 44 Byte per record



Confidence threshold: 1

Knowledge threshold:

- (ℓ,k,m)=(0,1000,0)
- (ℓ,k,m)=(3,1000,10)

4.4 GB

Conclusion and Future Work

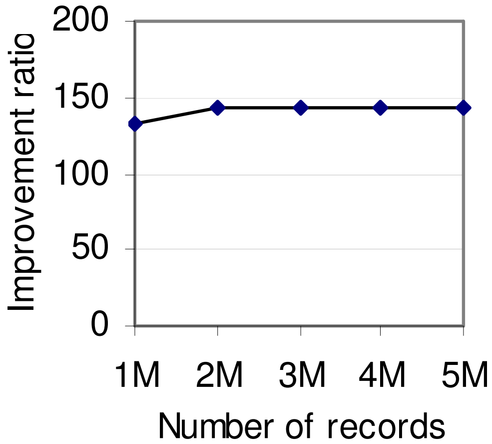
- It is important to consider adversarial knowledge in data privacy
- Tradeoff between expressiveness and feasibility
  - Useful expressions that satisfy the congregation property
- Future directions:
  - Other kinds of adversarial knowledge
    - Probabilistic knowledge expressions
    - knowledge about various kinds of social relationships
  - Other kinds of data
    - Search logs
    - Social networks

Thank You!

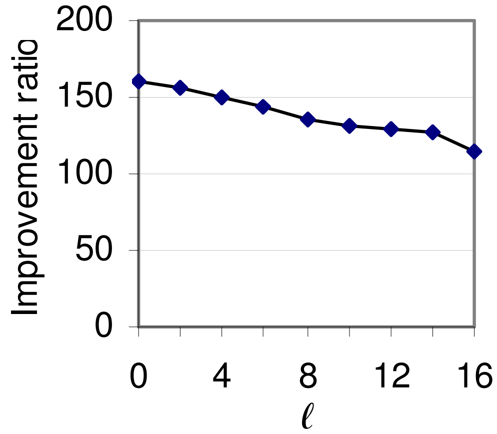
Supplementary Slides

Efficiency of SkylineCheck

(ℓ=10, k =10, m =10)



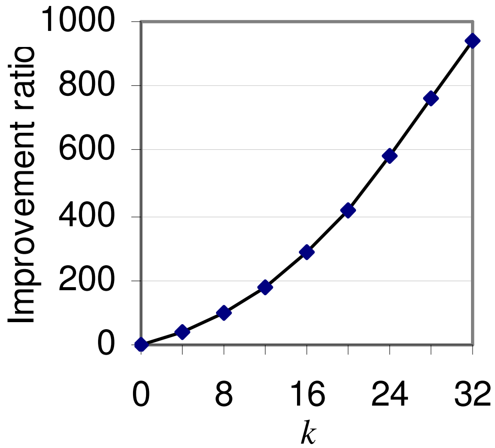
(ℓ: x-axis, k =10, m =10)



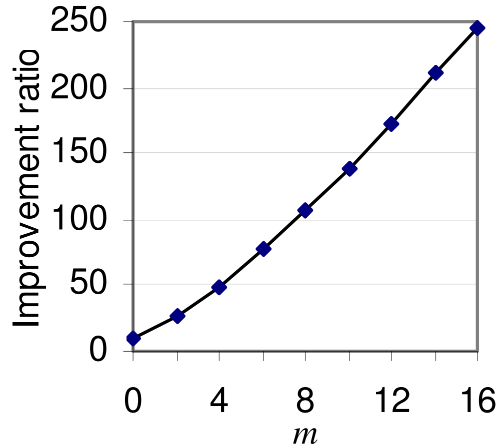
Improvement ratio =

$$\frac{\text{Execution time of DP}}{\text{Execution time of ours}}$$

(ℓ=10, k : x-axis, m =10)



(ℓ=10, k =10, m : x-axis)



Case Study: ℓ-Diverse Dataset

- Dataset: UCI adult dataset
  - Size: 45,222 records
  - Sensitive attribute: Occupation
- Create a (c=3, ℓ=6)-diverse release candidate $\mathbf{D}^*$
- How safe $\mathbf{D}^*$ is at confidence 0.95?
  - $\mathbf{D}^*$ is only safe for an adversary with knowledge beneath the knowledge skyline
  - E.g., if the adversary knows 5 people's occupations, then he can predict somebody t's occupation with confidence ≥ 0.95

Knowledge skyline of $\mathbf{D}^*$

ℓ k m
(0, 4, 0)
(1, 3, 1)
(2, 2, 2)
(3, 1, 2)
(2, 1, 3)
(4, 0, 3)
(3, 0, 4)

Related Work

- k-Anonymity (by Sweeney)
  - Each QI-group has at least k people
  - k-Anonymity is a special case of our 3D privacy criterion with knowledge $(0, k–2, 0)$ and confidence 1
    - Give each person a unique sensitive value
- ℓ-Diversity (by Machanavajjhala et al.)
  - Each QI-group has ℓ well-represented sensitive values
  - (c,ℓ)-Diversity is a special case of our 3D privacy criterion with knowledge $(\ell–2, 0, 0)$ and confidence $c/(c+1)$

Related Work

- Differential privacy & indistinguishability (Dwork et al.)
  - Add noise to query outputs so that no one can tell whether a record is in the original dataset with a high probability
- Probabilistic disclosure without adversarial knowledge
  - Xiao and Tao (SIGMOD’06 and VLDB’06)
  - Li et al. (ICDE’07)

Related Work

- Query-view privacy
  - Require complete independence between sensitive information and the released dataset
    - Deutsch et al. (ICDT'05), Miklau and Suciu (SIGMOD'04), and Machanavajjhala and Gehrke (PODS'06)
  - Bound the asymptotic probability of the answer of a Boolean query given views when the domain size $\rightarrow \infty$
    - Dalvi et al. (ICDT'05)

NP-Hardness

- $\max \Pr(t[S] = s \mid \mathbf{D}^*, K) < c$
  - $K = (A_1[S] = C_1 \leftrightarrow B_1[S] = D_1) \wedge \ldots \wedge (A_m[S] = C_m \leftrightarrow B_m[S] = D_m)$
    - $A_1, \ldots, A_m, B_1, \ldots, B_m, C_1, \ldots, C_m, D_1, \ldots, D_m$ are constants