Bottom-up Cell Suppression that Preserves the Missing-at-random Condition

Yoshitaka Kameya and Kentaro Hayashi
Meijo University

Outline

- Background
- Our proposal
- Experiments

Outline

- Background
  - Privacy-preserving data publishing
  - Bottom-up cell suppression
  - Incomplete data analysis
- Our proposal
- Experiments

Outline

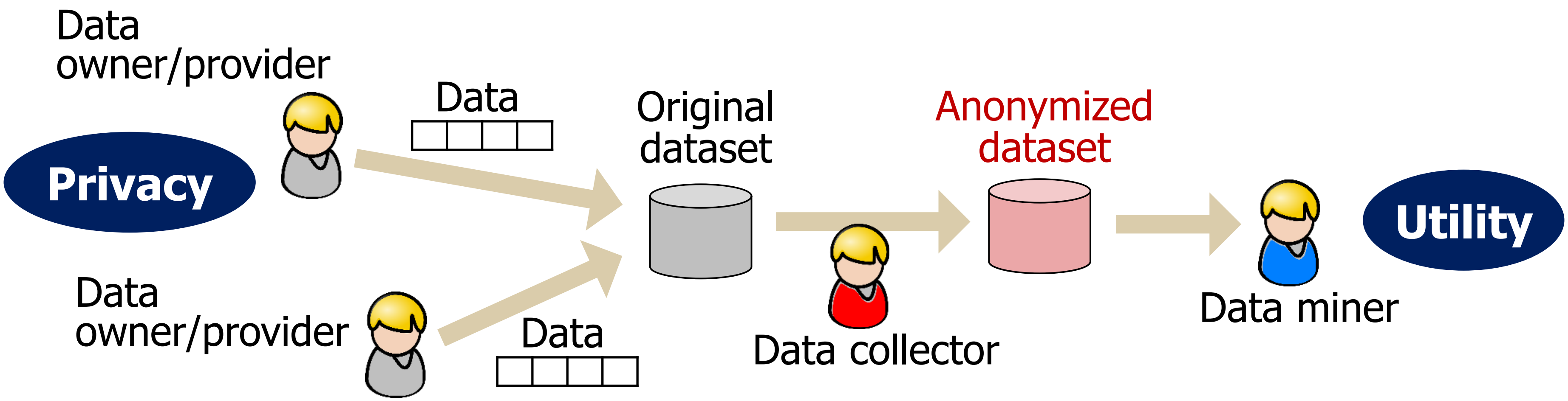
- Background
  - Privacy-preserving data publishing
  - Bottom-up cell suppression
  - Incomplete data analysis
- Our proposal
- Experiments

Privacy-preserving data publishing (1)

- In data mining: Fine-grained datasets ➔ Useful results
- Fine-grained *human-related* datasets
  ➔ Re-identification of a person
  ➔ Disclosure of his/her privacy
- Re-identification is possible easily by a combination of quasi-identifiers or QIDs (age, gender, etc.)

Privacy-preserving data publishing (2)

- Anonymization: Suppressing or generalizing (a part of) quasi-identifiers
- Privacy-preserving data publishing:
  - Needs to balance between **privacy** and **utility**



Privacy-preserving data publishing (3)

- k-anonymity:
  - Well-known privacy requirement
  - "Every tuple is not distinguishable from at least $k-1$ other tuples regarding QIDs"

2-anonymous dataset: (k = 2)

QIDs: Age, WorkClass, Gender; Sensitive attribute: Income

	Age	WorkClass	Gender	Income
2	[20, 30)	Government	Female	≤50K
	[20, 30)	Government	Female	≤50K
2	[20, 30)	Unemployed	Male	≤50K
	[20, 30)	Unemployed	Male	≤50K
2	[30, 40)	Private	Male	≤50K
	[30, 40)	Private	Male	≤50K
3	[30, 40)	Self-employed	Female	>50K
	[30, 40)	Self-employed	Female	≤50K
	[30, 40)	Self-employed	Female	>50K
2	[40, 50)	Government	Female	≤50K
	[40, 50)	Government	Female	≤50K

Probability of re-identification is at most $1 / k$ = 1/2

Outline

- Background
  - ✓Privacy-preserving data publishing
  - Bottom-up cell suppression
  - Incomplete data analysis
- Our proposal
- Experiments

Bottom-up cell suppression (1)

- Suppression
  - Often used in local recoding

Age	Nationality	Gender	Income
[20, 25)	Japan	Female	≤50K

→

Age	Nationality	Gender	Income
[20, 25)	Japan	**?**	≤50K

- Generalization
  - Often used in global recoding

Age	Nationality	Gender	Income
[20, 25)	Japan	Female	≤50K

→

Age	Nationality	Gender	Income
[20, 25)	**Asia**	Female	≤50K

- We focus on cell-suppresion:
  - Suppression does not require hierarchical knowledge
  - We have well-developed statistical tools (e.g. classifiers) that can handle suppressed values (*missing* values)

Bottom-up cell suppression (2)

- Rough pseudo code:

```
function Anonymize (k, D)
while there exists some tuple violating k-anonymity
    Pick up t violating k-anonymity
    t* := argmin_t' Γ(t, t', D);
    u := Suppress(t, t*);
    Update D by replacing t and t* with u
end;
return D;
```

Bottom-up cell suppression (2)

- Rough pseudo code:

k: the anonymity to achieve
D: the original dataset

```
function Anonymize (k, D)
while there exists some tuple violating k-anonymity
    Pick up t violating k-anonymity
    t* := argmin_t' Γ(t, t', D);
    u := Suppress(t, t*);
    Update D by replacing t and t* with u
end;
return D;
```

Bottom-up cell suppression (2)

- Rough pseudo code:

Repeatedly pick up at random a tuple violating k-anonymity

function Anonymize (k, D)

while there exists some tuple violating k-anonymity

  Pick up t violating k-anonymity

  $t^* := \mathrm{argmin}_{t'}\, \Gamma(t, t', D)$;

  $u := \mathrm{Suppress}(t, t^*)$;

  Update D by replacing t and t^* with u

end;

return D;

Bottom-up cell suppression (2)

- Rough pseudo code:

function Anonymize (k, D)
while there exists some tuple violating k-anonymity
  Pick up t violating k-anonymity
  $t^* := \mathrm{argmin}_{t'}\ \Gamma(t, t', D)$;
  $u := \mathrm{Suppress}(t, t^*)$;
  Update D by replac
end;
return D;

Suppression:
Create a new tuple where distinct QIDs between two tuples are suppressed

t

Age	Nationality	Gender	Income
[20, 25)	Japan	Female	≤50K

t^*

Age	Nationality	Gender	Income
[30, 35)	Japan	Male	≤50K

u

Age	Nationality	Gender	Income
?	Japan	?	≤50K

Γ: Suppression cost

Bottom-up cell suppression (2)

- Rough pseudo code:

```
function Anonymize (k, D)
  while there exists some t
      Pick up t violating k-anonym
      t* := argmin_t' Γ(t, t', D);
      u := Suppress(t, t*);
      Update D by replacing t and t* with u
  end;
  return D;
```

t^* is the counterpart of t such that:
- It belongs to t's class
- The suppression cost is minimum

Bottom-up cell suppression (2)

- Rough pseudo code:

```
function Anonymize (k, D)
  while there exists some tuple violating k-anonymity
      Pick up t violating k-anonymity
      t* := argmin_t' Γ(t, t', D);
      u := Suppress(t, t*);
      Update D by replacing t and t* with u
  end;
  return D;
```

Update the dataset:
Replace two old tuples with the new one

Bottom-up cell suppression (2)

- Rough pseudo code:

```
function Anonymize (k, D)
while there exists some tuple violating k-anonymity
    Pick up t violating k-anonymity
    t* := argmin_t' Γ(t, t', D);
    u := Suppress(t, t*);
    Update D by replacing t and t* with u
end;
return D;
```

Return k-anonymized dataset

Bottom-up cell suppression (3)

- Example

Original dataset

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	Private	Male	≤50K	1
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[30, 40)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	Government	Female	≤50K	1
[40, 50)	Government	Male	≤50K	1
[40, 50)	Unemployed	Female	≤50K	1

QIDs

Class label

# of duplicate tuples

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	Private	Male	≤50K	1
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[30, 40)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	Government	Female	≤50K	1
[40, 50)	Government	Male	≤50K	1
[40, 50)	Unemployed	Female	≤50K	1

Choose two tuples in the same class with the lowest suppression cost (Here we choose the closest two)

Bottom-up cell suppression (3)

- Example

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	Private	Male	≤50K	1
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[30, 40)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	?	Female	≤50K	2
[40, 50)	Government	Male	≤50K	1

Choose two again

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	Private	Male	≤50K	1
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[30, 40)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	Government	Female	≤50K	1
[40, 50)	Government	Male	≤50K	1
[40, 50)	Unemployed	Female	≤50K	1

Merge the chosen tuples with suppressing the conflicting values

Bottom-up cell suppression (3)

- Example

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	Private	Male	≤50K	1
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[30, 40)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	?	Female	≤50K	2
[40, 50)	Government	Male	≤50K	1

Suppress & Merge

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	?	Male	≤50K	2
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	?	Female	≤50K	2
[40, 50)	Government	Male	≤50K	1

Bottom-up cell suppression (3)

- Example

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
?	Government	Male	≤50K	**2**
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	?	Male	≤50K	2
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	?	Female	≤50K	2

Suppress & Merge

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	**1**
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	**?**	Male	≤50K	**2**
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	?	Female	≤50K	2
[40, 50)	Government	Male	≤50K	**1**

Bottom-up cell suppression (3)

- Example

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
?	Government	Male	≤50K	2
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	?	Male	≤50K	2
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	?	Female	≤50K	2

Age	WorkClass	Gender	Income	#
[20, 30)	?	Female	≤50K	2
?	Government	Male	≤50K	2
[20, 30)	Unemployed	?	≤50K	2
?	?	Male	≤50K	3
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	?	>50K	2
[40, 50)	?	Female	≤50K	2

These two tuples have
the same combination of QIDs

➔ Now the entire dataset has been
2-anonymized !

Bottom-up cell suppression (6)

- Example (summary)

Original dataset

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	Private	Male	≤50K	1
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[30, 40)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	Male	≤50K	1
[40, 50)	Self-employed	Male	>50K	1
[40, 50)	Government	Female	≤50K	1
[40, 50)	Government	Male	≤50K	1
[40, 50)	Unemployed	Female	≤50K	1

Anonymized dataset

Age	WorkClass	Gender	Income	#
[20, 30)	?	Female	≤50K	2
?	Government	Male	≤50K	2
[20, 30)	Unemployed	?	≤50K	2
?	?	Male	≤50K	3
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	?	>50K	2
[40, 50)	?	Female	≤50K	2

Utility:

How much information has been lost by anonymization?

Outline

- Background
    - ✓Privacy-preserving data publishing
    - ✓Bottom-up cell suppression
    - – Incomplete data analysis
- Our proposal
- Experiments

Incomplete data analysis (1)

- Target: Incomplete datasets (quite common in practice)
- Assumption:
  There is a *hidden* process making the complete dataset incomplete
- Many statistical tools have been developed assuming the missing-at-random (MAR) condition

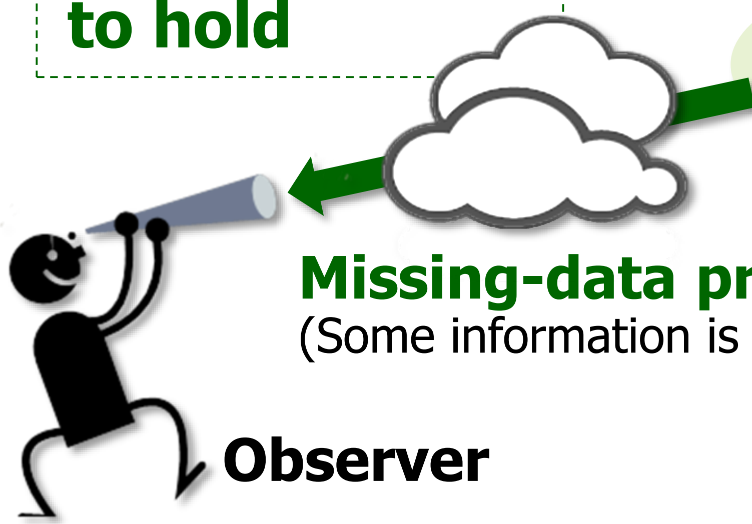
Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	Private	Male	≤50K	1
[30, 40)	Self-employed	Female	≤50K	1
[illegible]	[illegible]	Female	>50K	1
[illegible]	[illegible]	Male	≤50K	1
[illegible]	[illegible]	[illegible]	>50K	1
[illegible]	[illegible]	[illegible]	≤50K	1
[illegible]	[illegible]	Male	>50K	1
[illegible]	[illegible]	Female	≤50K	1
[40, 50)	Government	Male	≤50K	1
[40, 50)	Unemployed	Female	≤50K	1

Complete data

MAR assumed to hold

Age	WorkClass	Gender	Income	#
[20, 30)	?	Female	≤50K	2
?	Government	Male	≤50K	2
[20, 30)	Unemployed	?	≤50K	2
?	?	Male	≤50K	3
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	?	>50K	2
[40, 50)	?	Female	≤50K	2

Incomplete data



Missing-data process
(Some information is suppressed *by nature*)

Observer

Incomplete data analysis (2)

- **Key observation**: Anonymization process is an *artificial* process making the privacy dataset incomplete

→ We anonymize the dataset so that it satisfies MAR

→ The use of existing statistical tools will be safe
(They work as if the anonymization process never existed)

Age	WorkClass	Gender	Income	#
[20, 30)	?	Female	≤50K	2
?	Government	Male	≤50K	2
[20, 30)	Unemployed	?	≤50K	2
?	?	Male	≤50K	3
[30, 40)	Self-employed	Female	≤50K	1
[30, 40)	Self-employed	Female	>50K	1
[40, 50)	Self-employed	?	>50K	2
[40, 50)	?	Female	≤50K	2

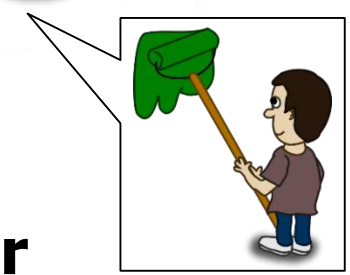
Anonymized dataset
(*Incomplete* data)

MAR designed to hold

Age	WorkClass	Gender	Income	#
[20, 30)	Private	Female	≤50K	1
[20, 30)	Government	Female	≤50K	1
[20, 30)	Government	Male	≤50K	1
[20, 30)	Unemployed	Female	≤50K	1
[20, 30)	Unemployed	Male	≤50K	1
[30, 40)	Private	Male	≤50K	1
[illegible]	[illegible]	Female	≤50K	1
[illegible]	[illegible]	[illegible]	>50K	1
[illegible]	[illegible]	[illegible]	≤50K	1
[illegible]	[illegible]	[illegible]	>50K	1
[illegible]	[illegible]	[illegible]	≤50K	1
[illegible]	[illegible]	[illegible]	>50K	1
[illegible]	[illegible]	Female	≤50K	1
[40, 50)	Government	Male	≤50K	1
[40, 50)	Unemployed	Female	≤50K	1

Dataset with privacy information
(*Complete* data)

Data user



Anonymization
(We *artificially* suppress some information)

Our goal

- We propose a cell-suppression based method for k-anonymization
  - Uses the notion from incomplete data analysis esp. the MAR condition
  - Justifies the use of Kullback-Leibler (KL) divergence [Kifer+ 06] as a utility measure
  - Incorporates KL divergence into a cell-suppression cost Γ in an efficient manner

Outline

✓ Background
- Our proposal
  - Naive Bayes
  - Missing-at-random condition
  - Kullback-Leibler divergence
- Experiments

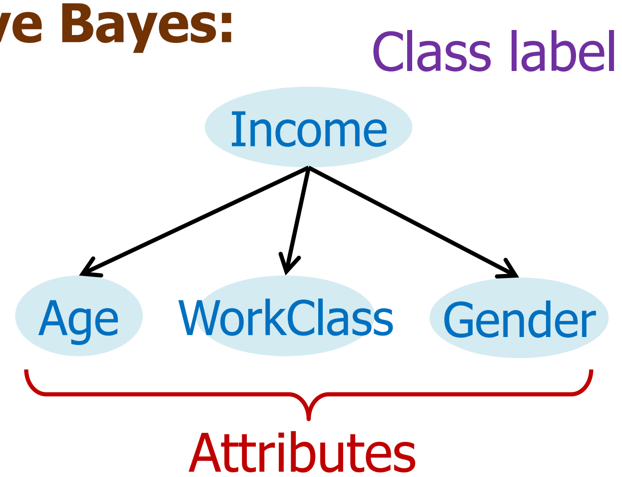
Proposed method: Naive Bayes (1)

- We focus on classification datasets
  (though the proposed method can handle non-classification dataset)
- Naive Bayes:
  – Assumes independence among attributes given a class label
  – Shows a good classification performance
    despite its simplicity

Attributes | Class label

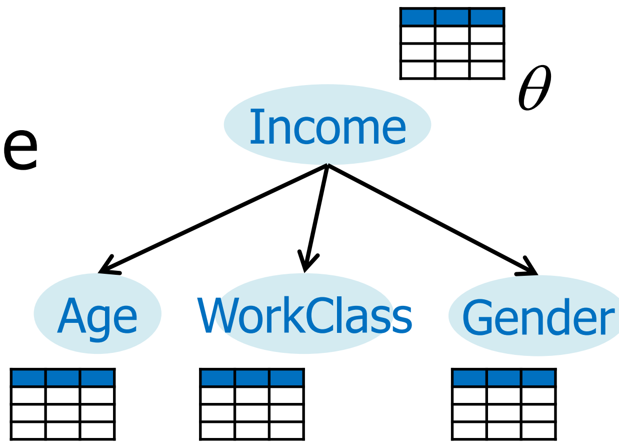
Age	WorkClass	Gender	Income
[20, 30)	Government	Female	≤50K
[20, 30)	Government	Female	≤50K
[20, 30)	Unemployed	Male	≤50K
[20, 30)	Unemployed	Male	≤50K
[30, 40)	Private	Male	≤50K
[30, 40)	Private	Male	≤50K
[30, 40)	Self-employed	Female	>50K
[30, 40)	Self-employed	Female	≤50K
[30, 40)	Self-employed	Female	>50K
[40, 50)	Government	Female	≤50K
[40, 50)	Government	Female	≤50K

Naive Bayes:



Proposed method: Naive Bayes (2)

- Naive Bayes's parameters θ:
  Entries in conditional probability table



- Learning θ in Naive Bayes:
  - Given a training dataset $D = \{t_1, t_2, \ldots, t_N\}$
  - Find $\theta*$ that maximize the likelihood:

$$\theta^* = \text{argmax}_\theta \prod_i p(t_i \mid \theta)$$

This learning scheme is called Maximum likelihood estimation (**MLE**)

- Prediction by the learned θ:
  - Given a new tuple $(x_1, x_2, \ldots, x_M)$ whose class label is unknown
  - Find the most probable class label c^* based on the current θ

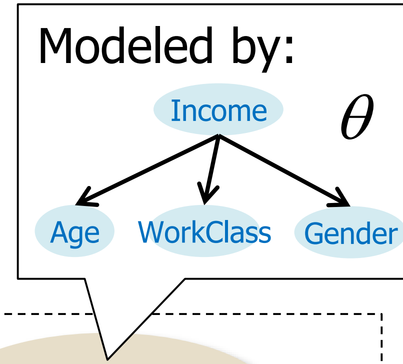
$$c^* = \text{argmax}_c \, p(c \mid \theta) \prod_j p(x_j \mid c, \theta)$$

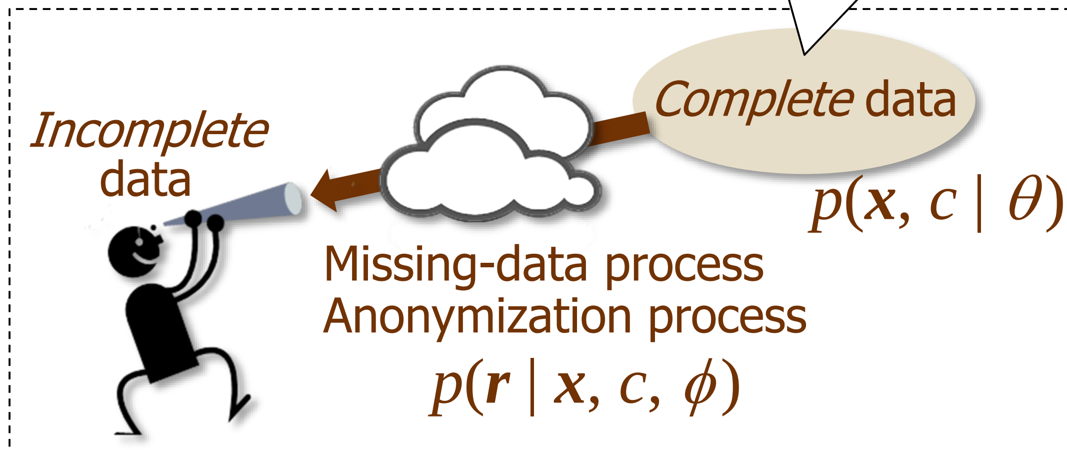
Proposed method: The MAR condition (1)

- Missing-data process with Naive Bayes:

$$\underbrace{p(\mathbf{r}, \mathbf{x}, c \mid \theta, \phi)}_{\text{Entire process}} = \underbrace{p(\mathbf{r} \mid \mathbf{x}, c, \phi)}_{\text{Missing-data process}} \; \underbrace{p(\mathbf{x}, c \mid \theta)}_{\text{Complete-data process}}$$

Missing-data indicator (Missingness)





- The MAR condition:
  Missingness of a cell-value does not depend on the value itself

$$\forall \mathbf{x}, c: p(\mathbf{r} \mid \mathbf{x}, c, \phi) = p(\mathbf{r} \mid \mathbf{x}_{\mathrm{obs}}, \mathbf{x}_{\mathrm{mis}}, c, \phi) = \underbrace{p(\mathbf{r} \mid \mathbf{x}_{\mathrm{obs}}, c, \phi)}_{\text{Missingness only depends on the non-suppressed part}}$$

Proposed method: The MAR condition (2)

- Under MAR, it is shown to be *safe* to learn θ based on the anonymized dataset

- We transform MAR into a more intuitive form:

MAR: $\forall \boldsymbol{x}, c: p(\boldsymbol{r} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}, c, \phi) = p(\boldsymbol{r} \mid \boldsymbol{x}_{\text{obs}}, c, \phi)$

$$\Rightarrow p(x_j \mid r_j = 0, c, \phi) = p(x_j \mid c, \phi)$$

$$\Leftrightarrow p(x_j \mid r_j = 1, c, \phi) = p(x_j \mid c, \phi)$$

Suppressed part must follow the original distribution

Non-suppressed part must follow the original distribution

We use KL divergence as a utility measure in anonymization

Kullback-Leibler (KL) **divergence** [Kifer+ 06] can be used to measure the deviation from MAR

Proposed method: KL divergence

- KL divergence: Dissimilarity between two distributions

$$\mathrm{KL}(\hat{p}, \hat{q}) = \sum_{\boldsymbol{x}, c} \hat{p}(\boldsymbol{x}, c) \log \frac{\hat{p}(\boldsymbol{x}, c)}{\hat{q}(\boldsymbol{x}, c)} = \sum_{c} \hat{p}(c) \sum_{j} \sum_{x_j} \hat{p}(x_j \mid c) \log \frac{\hat{p}(x_j \mid c)}{\hat{q}(x_j \mid c)}$$

$\hat{p}$: Distribution from the **original** dataset

$\hat{q}$: Distribution from the **anonymized** dataset (non-suppressed part of the original dataset)

- Difference between KL divergence *before* suppression and the one *after* suppression

$$\Delta\mathrm{KL} = \mathrm{KL}(\hat{p}, \hat{q}') - \mathrm{KL}(\hat{p}, \hat{q})$$

$\hat{p}$: Distribution from the **original** dataset

$\hat{q}$: Distribution from the **anonymized** dataset **before** suppression

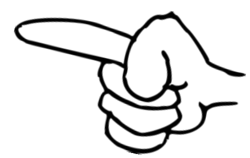
$\hat{q}'$: Distribution from the **anonymized** dataset **after** suppression

- $\Delta\mathrm{KL}$ is finally used as the cell-suppression cost Γ_{mar}

Proposed method: Summary

- We introduced a cost function Γ_{mar} which considers the MAR condition and KL divergence
- We plugged Γ_{mar} into a bottom-up cell-supression procedure:

function Anonymize (k, D)
while there exists some tuple violating k-anonymity
    Pick up t violating k-anonymity
    $t^* := \text{argmin}_{t'}\ \Gamma_{\text{mar}}(t, t', D)$;
    $u := \text{Suppress}(t, t^*)$;
    Update D by replacing t and t^* with u
end;
return D;



Outline

- ✓ Background
- ✓ Our proposal
  - ✓ Naive Bayes
  - ✓ Missing-at-random condition
  - ✓ Kullback-Leibler divergence
- Experiments

Experiments: Settings (1)

- **Target**: the Adult dataset from UCI ML Repository
- We measured the degree of utility loss under the costs:
  - Γ_{ham} (ham): Based on Hamming distance
    ➔ Minimize the number of suppressions

    No consideration on probability distribution
  - Γ_{info} (info): Based on self-information [Harada+ 12]
    ➔ Suppress frequent values first

    Considering local (individual) probabilities
  - Γ_{mar} (mar): Based on the missing-at-random (MAR) condition and KL divergence (our proposal)

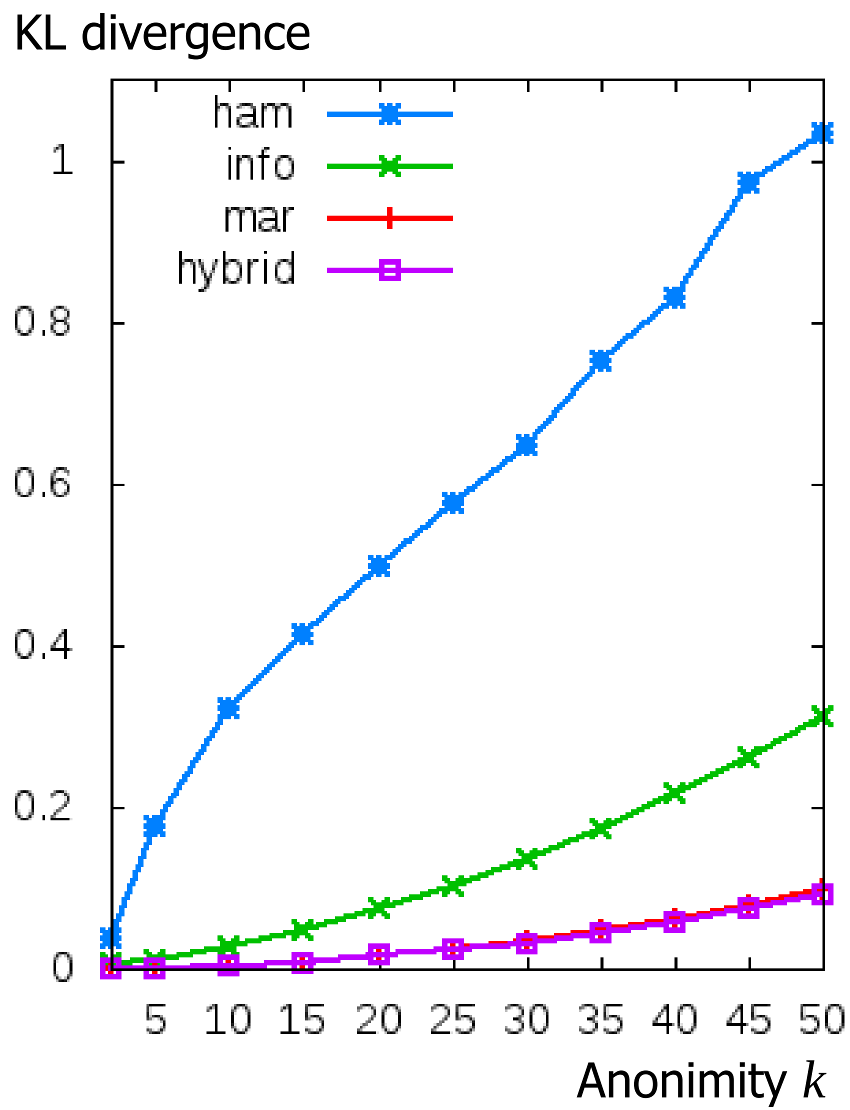
    Considering the entire distribution
  - Γ_{hybrid} (hybrid): A simple hybrid of Γ_{ham} and Γ_{mar}

Experiments: Settings (2)

- Utility loss is measured by:
  - KL divergence
  - Error rate in classification
    (under stratified 10-fold cross-validation)
- Classifiers implemented in Weka:
  - Naive Bayes (primary)
  - C4.5
- Preprocessing:
  - Picked up 8 QIDs also used in previous work
    (Age, Work class, Education, Marital status, Occupation, Race, Gender, Native country)
  - Discretized the Age attribute

Experiments: KL divergence

- Anonymity k was varied from 2 to 50
- Γ_{mar} and Γ_{hybrid} achieved quite small degradation as expected
- Γ_{ham} worked worst since it does not consider probability distribution
- Γ_{info} was moderate

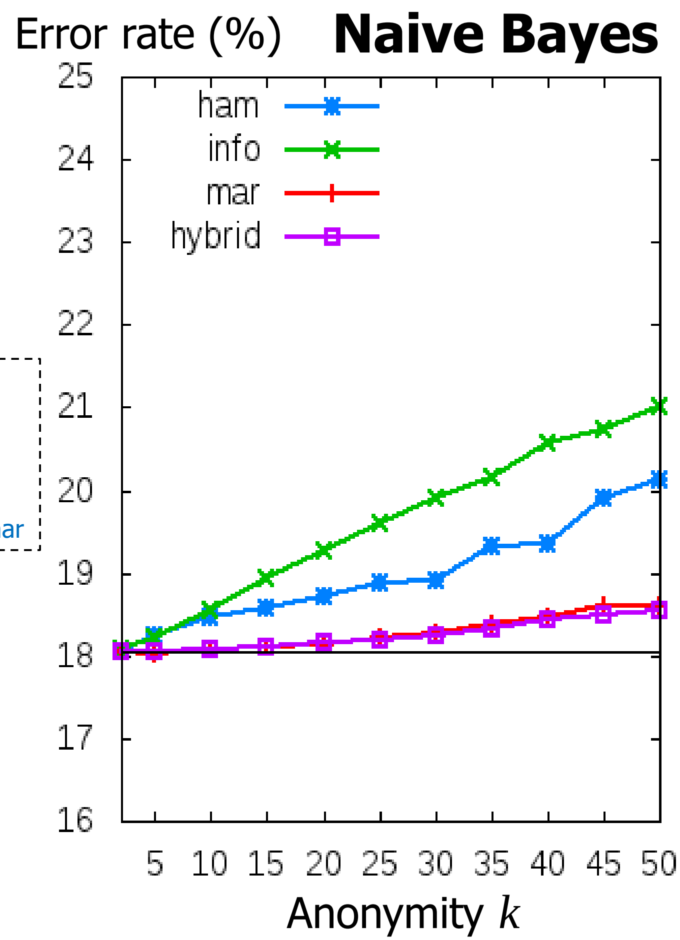


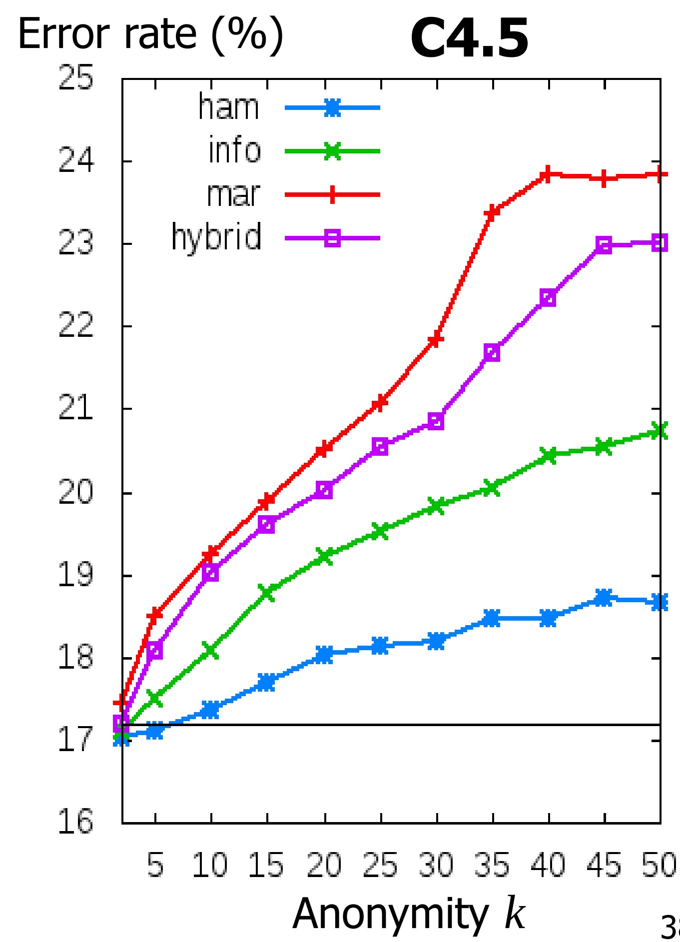
Γ_{ham}: Hamming distance
Γ_{info}: Self-information
Γ_{mar}: Our proposal
Γ_{hybrid}: Hybrid of Γ_{ham} and Γ_{mar}

Experiments: Classification performance

- Naive Bayes worked better with Γ_{mar} and Γ_{hybrid} as expected
- C4.5 worked best with Γ_{ham}
  (C4.5 seems *not* to be robust against missing values)

Γ_{ham}: Hamming distance
Γ_{info}: Self-information
Γ_{mar}: Our proposal
Γ_{hybrid}: Hybrid of Γ_{ham} and Γ_{mar}

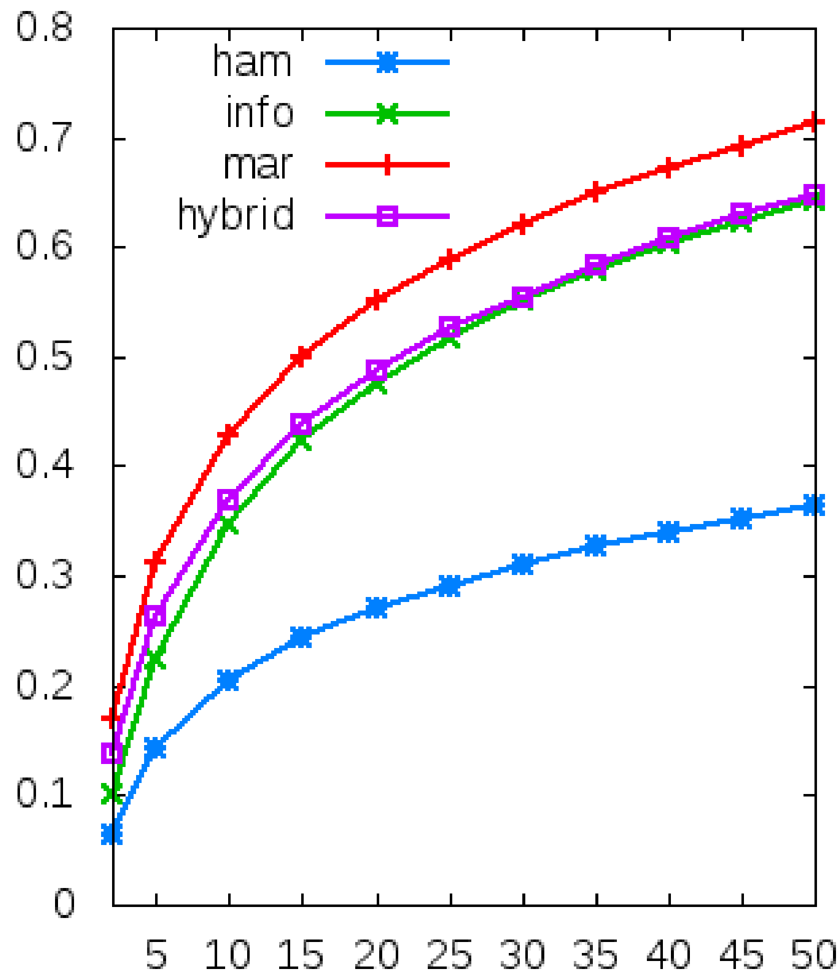




Experiments: Suppression ratio

- Opposite behaviors were observed
- Γ_{ham} keeps the smallest the number of suppressed cells
- Γ_{mar} tends to perform many suppressions
- Γ_{info} and Γ_{hybrid} were moderate

Suppression ratio (ranges from 0 to 1)



Anonimity k

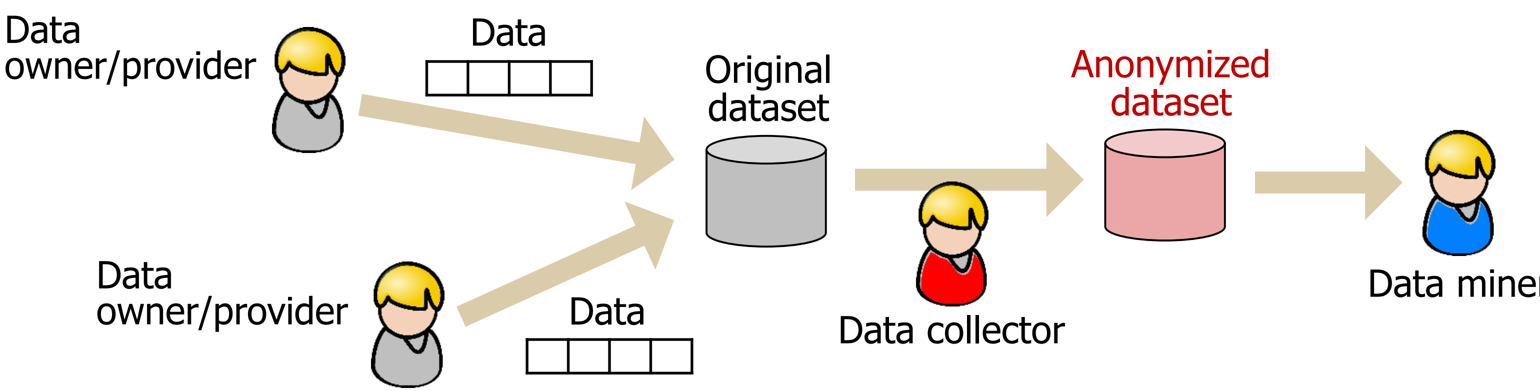
Γ_{ham}: Hamming distance
Γ_{info}: Self-information
Γ_{mar}: Our proposal
Γ_{hybrid}: Hybrid of Γ_{ham} and Γ_{mar}

Summary

- We proposed a new cell-suppression based method for k-anonymization:
  - Uses the notion from incomplete data analysis esp. the MAR condition
  - Justifies the use of Kullback-Leibler (KL) divergence as a utility measure
  - Incorporates KL divergence into a cell-suppression cost in an efficient manner
  - Worked as expected for a benchmark dataset

Open problems

- Removal of the independence assumption in naive Bayes
- Multi-objective optimization
  - Introducing a classification-centric measure
  - Considering l-diversity [Machanavajjhala+ 07]
  - Different roles in privacy-preserving data publishing



- Cell-generalization using hierarchical knowledge
  - The coarsening-at-random condition [Heitjan+ 91]