

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



# Coursework : Privacy preserving salary prediction in data mining

GV: PGS.TS ĐẶNG TRẦN KHÁNH

---

Chu Xuân Tình -1870583  
Nguyễn Đức Huy -1870567

# Coursework

---

- ❑ Introduction
- ❑ Data input & Model
- ❑ Data anonymization tool
- ❑ Implementation

# Introduction

---

## ❑ Privacy preserving salary prediction in data mining

## ❑ Data input

A	B	D	E	F	G	H	I	J	K	M	N	O
age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	hours-per-w	native-country	class
39	State-gov	Bachelors		13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	40	United-States <=50K.
50	Self-emp-nc	Bachelors		13	Married-civ-spouse	Exec-manag	Husband	White	Male	0	13	United-States <=50K.
38	Private	HS-grad		9	Divorced	Handlers-cle	Not-in-family	White	Male	0	40	United-States <=50K.
53	Private	11th		7	Married-civ-spouse	Handlers-cle	Husband	Black	Male	0	40	United-States <=50K.
28	Private	Bachelors		13	Married-civ-spouse	Prof-special	Wife	Black	Female	0	40	Cuba <=50K.
37	Private	Masters		14	Married-civ-spouse	Exec-manag	Wife	White	Female	0	40	United-States <=50K.
49	Private	9th		5	Married-spouse-absent	Other-servic	Not-in-family	Black	Female	0	16	Jamaica <=50K.
52	Self-emp-nc	HS-grad		9	Married-civ-spouse	Exec-manag	Husband	White	Male	0	45	United-States >50K.
31	Private	Masters		14	Never-married	Prof-special	Not-in-family	White	Female	14084	50	United-States >50K.
42	Private	Bachelors		13	Married-civ-spouse	Exec-manag	Husband	White	Male	5178	40	United-States >50K.
37	Private	Some-colleg		10	Married-civ-spouse	Exec-manag	Husband	Black	Male	0	80	United-States >50K.
30	State-gov	Bachelors		13	Married-civ-spouse	Prof-special	Husband	Asian-Pac-Is	Male	0	40	India >50K.
23	Private	Bachelors		13	Never-married	Adm-clerical	Own-child	White	Female	0	30	United-States <=50K.
32	Private	Assoc-acdm		12	Never-married	Sales	Not-in-family	Black	Male	0	50	United-States <=50K.
40	Private	Assoc-voc		11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Is	Male	0	40	? >50K.
34	Private	7th-8th		4	Married-civ-spouse	Transport-m	Husband	Amer-Indiar	Male	0	45	Mexico <=50K.
25	Self-emp-nc	HS-grad		9	Never-married	Farming-fis	Own-child	White	Male	0	35	United-States <=50K.
32	Private	HS-grad		9	Never-married	Machine-op	Unmarried	White	Male	0	40	United-States <=50K.
38	Private	11th		7	Married-civ-spouse	Sales	Husband	White	Male	0	50	United-States <=50K.
43	Self-emp-nc	Masters		14	Divorced	Exec-manag	Unmarried	White	Female	0	45	United-States >50K.

# Data anonymization tool

## □ Data anonymization tool – ARX Tool

The screenshot displays the ARX Anonymization Tool interface. The main window is titled "ARX Anonymization Tool - TINH2".

**Input data:** A table showing a sample of 32561 rows from the adult dataset. The columns include age, workclass, fnlwgt, education, education-num, marital-status, and others.

**Data transformation:** A panel on the right shows transformation rules for the "age" attribute. The type is set to "Quasi-identifying" and the transformation is "Generalization". The table lists age values grouped into ranges: 17-20, 21-25, 26-30, etc., with each range mapped to a single value marked with an asterisk (\*).

**Privacy models:** A section at the bottom left shows a 5-Anonymity model. It includes tabs for General settings, Utility measure, Coding model, and Attribute weights. Settings include a suppression limit of 0% and an approximate checkbox for "Assume practical monotonicity".

# Data anonymization tool

## □ Data anonymization tool – ARX Tool

The screenshot displays the ARX Anonymization Tool interface, specifically version TINH2. The window is divided into several sections:

- Top Bar:** ARX Anonymization Tool - TINH2, File, Edit, View, Help.
- Toolbar:** Includes icons for opening files, saving, zooming, and other common operations.
- Status Bar:** Attribute: fnlwgt Transformations: 5 Selected: [1, 0] Applied: [1, 0]
- Navigation:** Configure transformation, Explore results, Analyze utility, Analyze risk.
- Input Data Table:** Shows a list of 100 rows of data with columns: age, workclass, fnlwgt, education, education-num, marital-status. Rows 13340 through 13359 are visible.
- Output Data Table:** Shows the same 100 rows after anonymization. The output values are identical to the input values in this specific view.
- Bottom Panels:** Two panels for "Summary statistics", "Distribution", "Contingency", "Class sizes", "Properties", and "Classification models". Each panel has a "Parameter" table and a "Value" table.

# Data anonymization tool

- Data anonymization tool - <https://amnesia.openaire.eu/>

The screenshot shows the Amnesia data anonymization tool interface. On the left is a sidebar with navigation links: Anonymization Wizard, Restart, Source (Manage, Load From Local, Load From Zenodo), Anonymized, Hierarchy, Algorithms, Solution Graph, and Results. The main area displays a dataset titled "Adult2Training.csv". At the top of this area, there are buttons for "Load XML File", "Load New Dataset", "Save To Local", "Save To Zenodo", and "Load Anon Rules". Below these buttons, a table shows 10 entries from the dataset. The columns are: age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, and hours-per-week.

age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40
37	Private	284582	Masters	14	Married-civ-spouse	Exec-admin	Wife	White	Female	0	0	40

# Implementation

38 Private 215646 HS-grad 9 Divorced Handlers- Not-in-family White Male 0 0 40 United-States

Anonymized

Hierarchy

Algorithms

Solution Graph

Results

Amnesia Site

Check if dataset is Anonymous

Choose attributes :

- age
- workclass
- fnlwgt
- education
- education-num
- marital-status
- occupation
- relationship
- race
- sex
- capital-gain
- capital-loss
- hours-per-week
- native-country
- class

Choose K :

4

Close Show Anonymization Proceed to Hierarchies

© 2016, All Rights Reserved.

# Implementation

38 Private 215646 HS-grad 9 Divorced Handlers- Not-in-family White Male 0 0 40 United-States 40 United-States 40 Cuba 40 United-States 16 Jamaica 45 United-States 50 United-States 40 United-States

Anonymized

Hierarchy

Algorithms

Solution Graph

Results

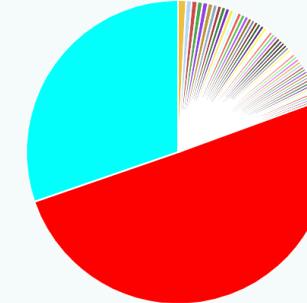
Amnesia Site

Check if dataset is Anonymous

"age,education,marital-status,sex,native-country" is anonymous for k : 1.

Percentage of displayed dataset is : 100%

To produce a k = 4 suppress 50.28%



Close Suppress Apply

Showing 1 to 10 of 3,998 entries

Previous 1 2 3 4 5 ... 400 Next

Check Anonymization Proceed to Hierarchies

© 2016, All Rights Reserved.

# Implementation

The screenshot shows the Amnesia software interface. On the left is a sidebar with the following menu items:

- Anonymization Wizard
- Restart
- Source
- Anonymized
- Hierarchy (selected, indicated by a green border)
- Manage
- Load from Local
- Auto Generate
- Algorithms
- Solution Graph
- Results
- Advanced

The main area is titled "Hierarchy Autogenerate Wizard" and "version:1.2.1 beta". It displays the "Hierarchy Autogenerate" step, with "2. Hierarchy Info" highlighted in a green box. The "Hierarchy Information" section contains the following fields:

Step	10
Name	age_mask
Domain start	0
End limit	100
Fanout	3

At the bottom right are buttons for "Previous", "Finish", and "Cancel". At the very bottom of the main window is a footer with the text "© 2016, All Rights Reserved."

# Implementation

Anonymization Wizard

Hierarchy Name : age\_mask

Edit

Manage

Load from Local

Auto Generate

Algorithms

Solution Graph

Results

Amnesia Site

(null)

0-100

60-100

30-60

0-30

50-60

40-50

30-40

Proceed to Algorithms

```
graph TD; 0_100[0-100] --> null((null)); 0_100 --> 60_100[60-100]; 0_100 --> 30_60[30-60]; 0_100 --> 0_30[0-30]; 30_60 --> 50_60[50-60]; 30_60 --> 40_50[40-50]; 30_60 --> 30_40[30-40]
```

© 2016, All Rights Reserved.

# Implementation

The screenshot shows the Amnesia software interface. On the left is a dark sidebar with the Amnesia logo at the top. Below the logo are several menu items: Anonymization Wizard, Restart, Source, Anonymized, Hierarchy (which is expanded to show Manage, Load from Local, and Auto Generate), Algorithms, Solution Graph, Results, and Help. The main area is titled "Hierarchy Autogenerate Wizard" and "version:1.2.1 beta". It displays the "Hierarchy Autogenerate" step, with "2. Hierarchy Info" highlighted in a green bar. The "Hierarchy Information" section contains fields for Sorting (set to random), Name (set to relation\_mask), and Fanout (set to 2). At the bottom right are buttons for Previous, Finish, and Cancel. A copyright notice at the bottom states "© 2016, All Rights Reserved."

Hierarchy Autogenerate Wizard  
version:1.2.1 beta

Hierarchy Autogenerate

1. Choose Attribute and Hierarchy Type    2. Hierarchy Info

Hierarchy Information

Sorting: random

Name: relation\_mask

Fanout: 2

Previous    Finish    Cancel

© 2016, All Rights Reserved.

# Implementation

- Anonymization Wizard
- Restart
- Source
- Anonymized
- Hierarchy
  - Manage
  - Load from Local
  - Auto Generate
- Algorithms
- Solution Graph
- Results
- Amnesia Site

### Hierarchy

Hierarchy Name : relation\_mask

Edit

```
graph TD; Random5((Random5)) --- null((null)); Random5 --- Random4((Random4)); Random5 --- Random3((Random3)); Random3 --- Random1((Random1)); Random3 --- Random0((Random0));
```

Own-child  
Not-in-family

Up, Down, Left, Right navigation icons are located at the bottom left. Minus, Plus, and Cross icons are located at the bottom right.

Proceed to Algorithms

# Implementation

The screenshot shows the Amnesia software interface. On the left is a dark sidebar with a circular logo containing a stylized 'E' and vertical bars, labeled 'amnesia'. The sidebar includes navigation links: Anonymization Wizard, Restart, Source, Anonymized, **Hierarchy** (which is expanded to show Manage, Load from Local, and Auto Generate), Algorithms, Solution Graph, and Results.

The main area is titled 'Hierarchy Autogenerate Wizard' with 'version:1.2.1 beta' below it. A progress bar at the top indicates '1. Choose Attribute and Hierarchy Type' and '2. Hierarchy Info' (the current step). The 'Hierarchy Information' section contains fields for 'Sorting' (set to 'alphabetical'), 'Name' (set to 'country\_mask'), and 'Fanout' (set to '5'). At the bottom right of this section are buttons for 'Previous', 'Finish', and 'Cancel'. The footer of the main window says '© 2016, All Rights Reserved.'

# Implementation

Anonymization Wizard

- Restart
- Source
- Anonymized
- Hierarchy
  - Manage
  - Load from Local
  - Auto Generate
- Algorithms
- Solution Graph
- Results
- Amnesia Site

Hierarchy Name : country\_mask

Add Node | Add Edge | X

```
graph TD; Random10 --> null((null)); Random10 --> asia((asia)); Random10 --> europ((europ)); europ --> Random4((Random4)); europ --> Random3((Random3)); europ --> Random2((Random2)); europ --> Random1((Random1)); europ --> Random0((Random0)); Random2 --> Greece((Greece)); Random2 --> Guatemala((Guatemala)); Random2 --> France((France)); Random2 --> Haiti((Haiti)); Random2 --> Germany((Germany))
```

Proceed to Algorithms

# Implementation

42	Private	159449	Bachelors	13	Married-civ-spouse	r
----	---------	--------	-----------	----	--------------------	---

Showing 1 to 10 of 3,998 entries

Previous 1 2 3 4 5  
... 400 Next

### Bind Hierarchies with Attributes

Indicate with generalization hierarchy will be used for each dataset attribute. The same hierarchy can be used in multiple attributes. A hierarchy must be defined for each quasi identifier.

age	age_mask
workclass	
fnlwgt	
education	
education-num	
marital-status	

### Algorithms

Type: Flash ▾ K: - 5 + Execute

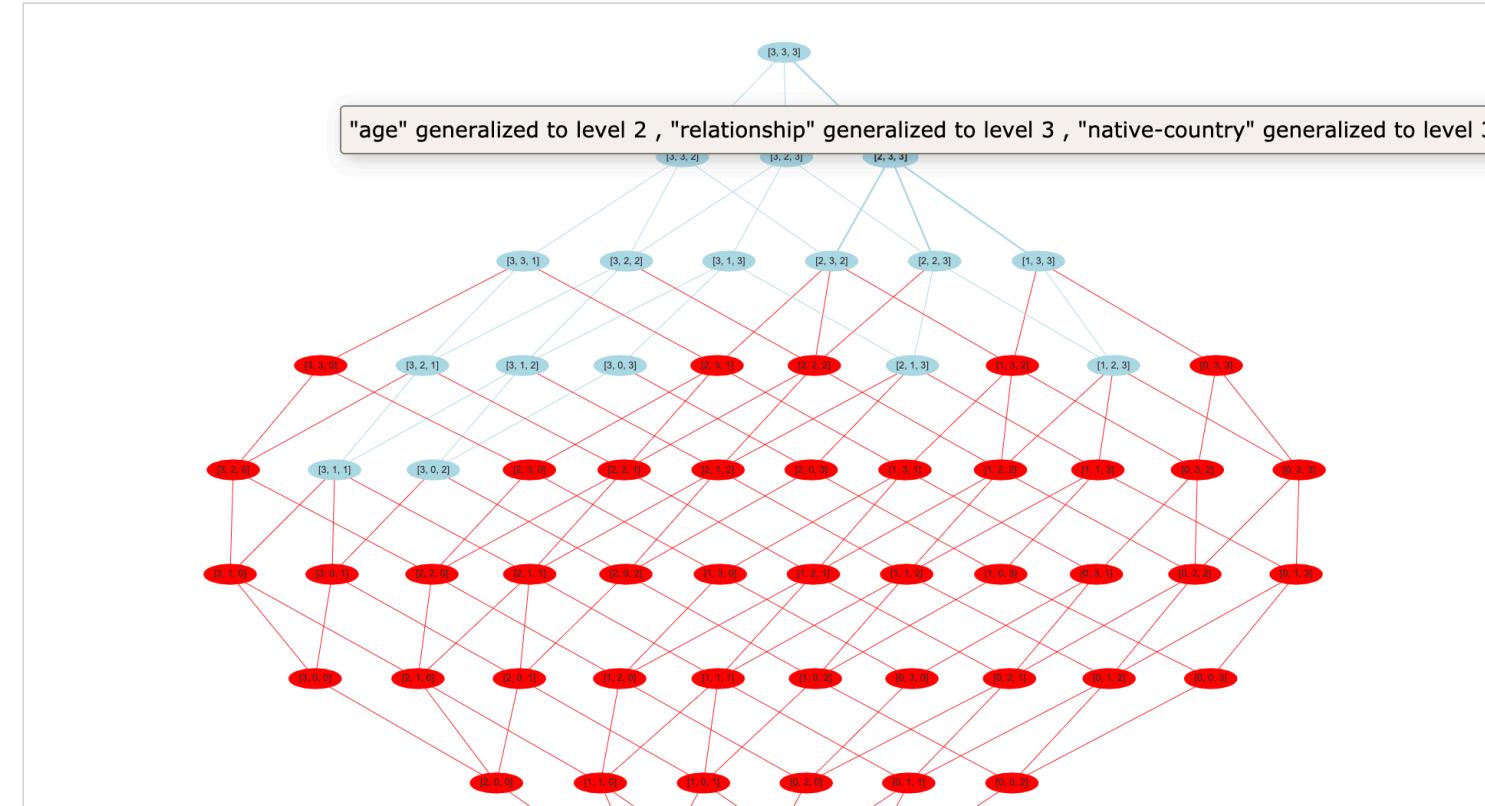
© 2016, All Rights Reserved.

# Implementation



version:1.2.1 beta

Explore the solution space. Blue nodes indicate safe solutions and red nodes unsafe. Hover a node to view the generalizations levels of the attributes. Select a node to view a sample of the anonymized dataset and to explore its statistical properties. Unsafe solutions can be transformed to safe by using suppression. Double-click a node to apply a solution.



# Implementation

Anonymization Wizard

Restart

Source

Anonymized

Hierarchy

Algorithms

**Solution Graph**

Results

Amnesia Site

[3, 3, 3]

### Statistics of the dataset with this solution

Choose Attributes : age

Percentage of displayed dataset is : 100%

A pie chart illustrating the distribution of the 'age' attribute. The chart is divided into three segments: a large blue segment (43%, [Random0]), a smaller yellow segment (56%, [Random1]), and a very small red segment (1%, [Random2]).

Generalized to level 1

Close   Suppress   Apply

© 2016, All Rights Reserved.

# Implementation

The screenshot shows the Amnesia software interface. On the left, a dark sidebar lists navigation options: Anonymization Wizard, Restart, Source, Anonymized, Hierarchy, Algorithms, Solution Graph (which is highlighted with a teal bar at the top), Results, and Amnesia Site. The main area is titled "Solutions Graph" and contains a table titled "Anonymized Dataset". The table has 12 columns: location, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, and class. The data consists of 8 rows, each representing a different individual with various demographic and socioeconomic details. A large red oval highlights the "native-country" column in the last row. To the right of the table, a vertical bar has the text "generalized to level 2" written vertically.

location	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	class
bachelors	13	Never-married	Adm-clerical	Random3	White	Male	2174	0	40	asia	<=50K.
bachelors	13	Married-civ-spouse	Exec-managerial	Random3	White	Male	0	0	13	asia	<=50K.
hs-grad	9	Divorced	Handlers-cleaners	Random3	White	Male	0	0	40	asia	<=50K.
th	7	Married-civ-spouse	Handlers-cleaners	Random3	Black	Male	0	0	40	asia	<=50K.
bachelors	13	Married-civ-spouse	Prof-specialty	Random4	Black	Female	0	0	40	europe	<=50K.
asters	14	Married-civ-spouse	Exec-managerial	Random4	White	Female	0	0	40	asia	<=50K.
h	5	Married-spouse-absent	Other-service	Random3	Black	Female	0	0	16	europe	<=50K.

---

THANK YOU!