Sanitizing and Minimizing Databases for Software Application Test Outsourcing

Boyang Li

College of William and Mary

Mark Grechanik

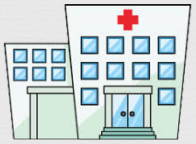
University of Illinois at Chicago

Denys Poshyvanyk

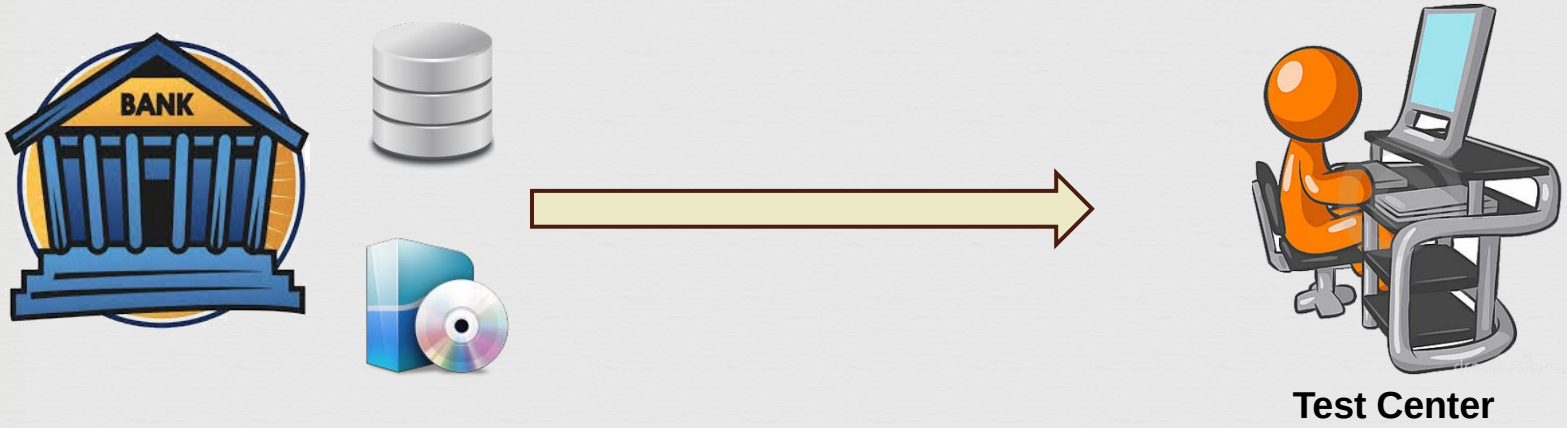
College of William and Mary

Motivation

- Database-Centric Applications (DCA)



Motivation



- Expected outsourcing market in 2020: $50B vs. $30B in 2010
- State of the art: clean room testing and fake data generation

Clean Room Testing

Generate Fake Data

Original table

Age	Gender	# of Children	Nationality
51	M	1	Chinese
29	F	3	American
46	M	1	Japanese
9	F	0	American
…	…	…	…

Anonymized table (Naïve version)

Age	Gender	# of Children	Nationality
61	M	3	Chinese
8	F	2	Chinese
33	F	0	Japanese
29	M	1	American
…	…	…	…

- Type and value restriction
- Semantic connections between data
- Program behavior

```
if (nationality == “Japanese” && age > 40) {
        f(disease);
}
```

Our work

- We focus on balancing the following four dimensions:

Privacy	**Data minimization**
Semantic correctness	**Testing coverage**

PISTIS - Protecting and mInimizing databases for Software TestIng taSks

- Program analysis
- Weight-based k-clustering algorithm
- Compute centroid objects
- Associative rule mining

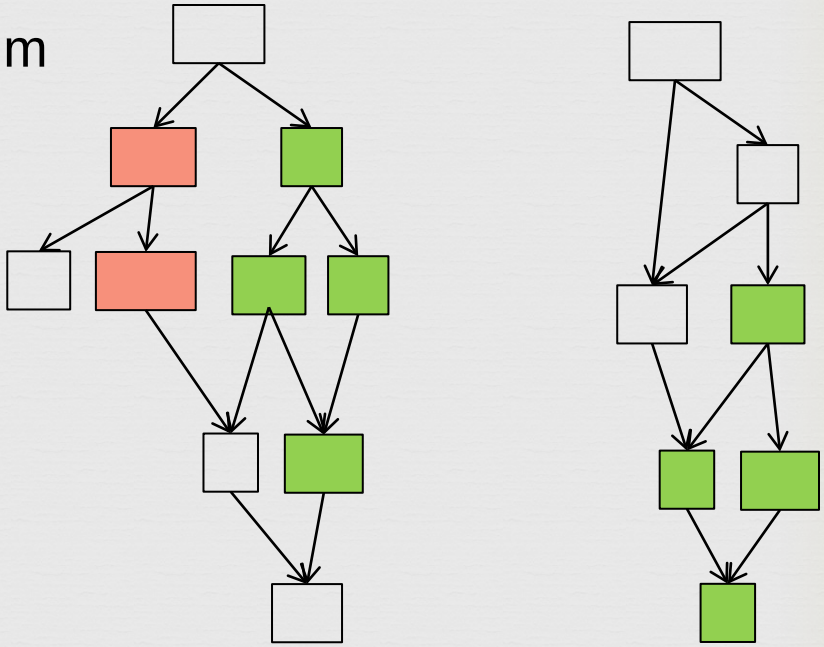
PISTIS - Program analysis

- Program analysis

Nationality	Age
…	…
…	…

Original table

- Weight-based k-clustering algorithm
- Compute centroid objects
- Associative rule mining

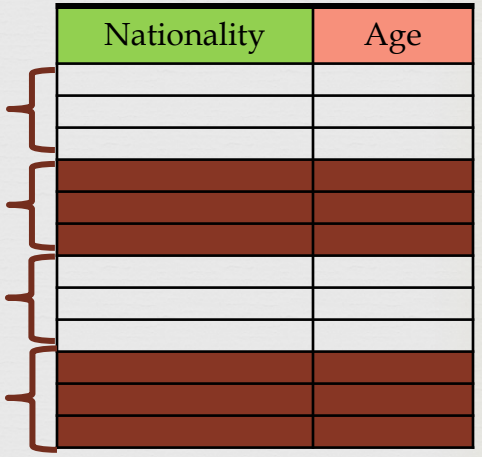


CFGs

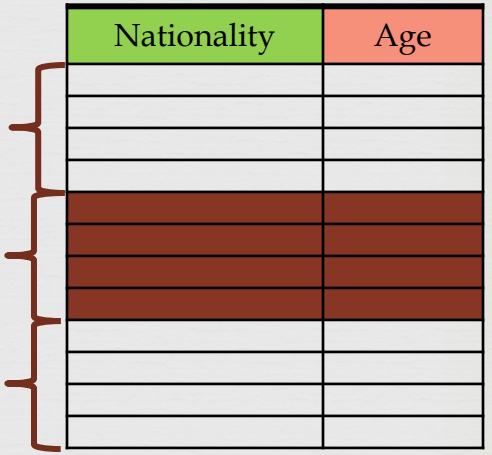
PISTIS - Weight-based k-clustering algorithm

- Program analysis
- Weight-based k-clustering algorithm
- Compute centroid objects
- Associative rule mining

K=4

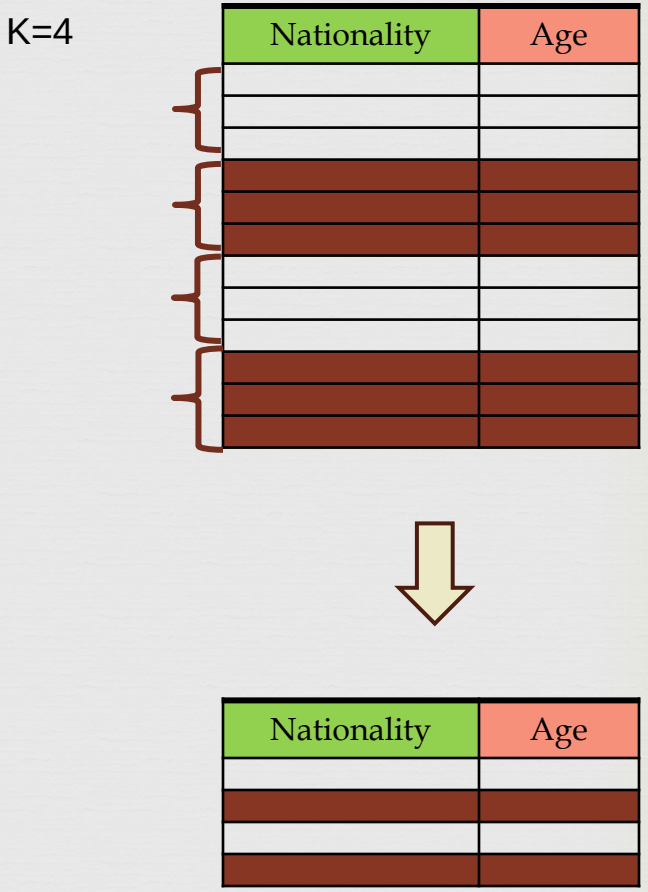


K=3



PISTIS - Compute centroid objects

- Program analysis
- Weight-based k-clustering algorithm
- Compute centroid objects
- Associative rule mining



PISTIS - Compute centroid objects

- Program analysis
- Weight-based k-clustering algorithm
- Compute centroid objects
- Associative rule mining

Original table

“pregnant ==true -> gender == female”

PISTIS - Protecting and mInimizing databases for Software TestIng taSks

- Program analysis – – – – – – – – **Test coverage**
- Weight-based k-clustering algorithm
- Compute centroid objects

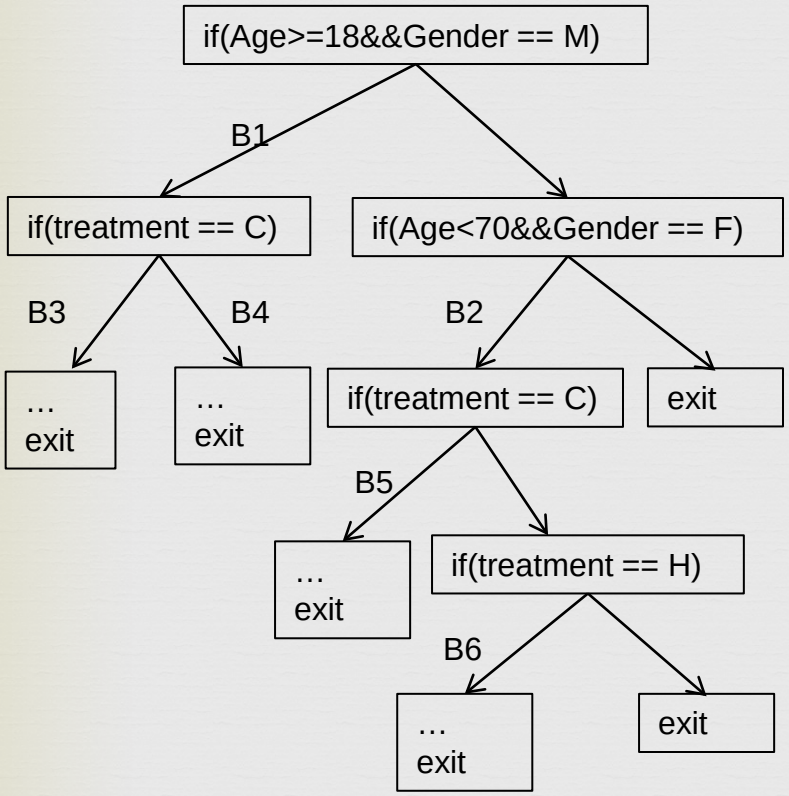
Data minimization

Privacy

- Associative rule mining – – – – – – – – **Semantic correctness**

PISTIS allows us to meet all four goals

An example

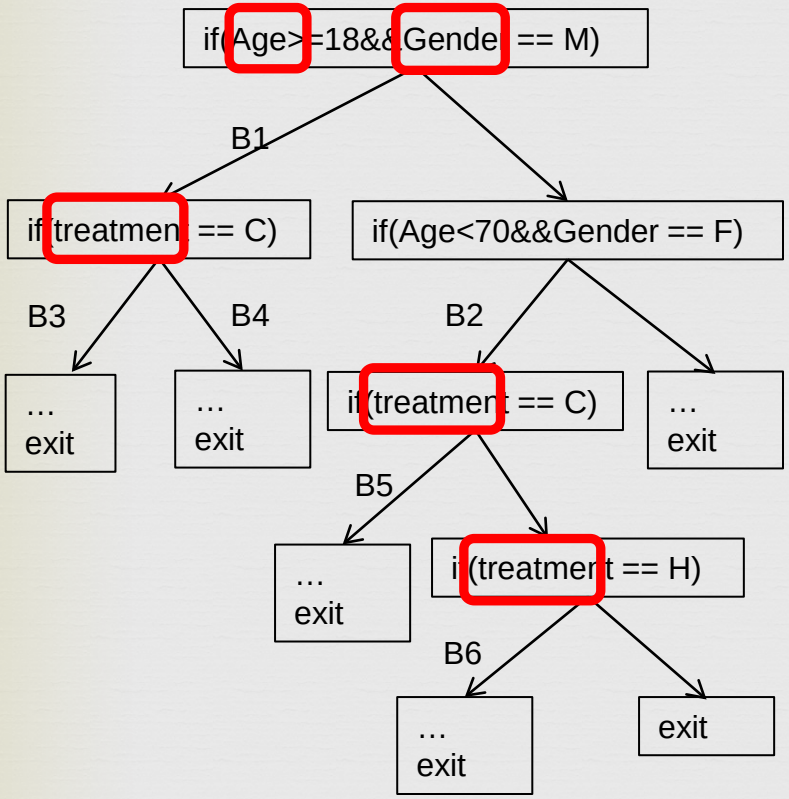


Original table

Age	State	Gender	Treatment	Branch
42	OH	M	Vasectomy	B1,B4
47	OH	F	Hysterectomy	B2,B6
51	VA	F	Chemotherapy	B2,B5
55	VA	M	Chemotherapy	B1,B3
62	OH	M	Chemotherapy	B1,B3
67	CA	F	Hysterectomy	B2,B6
30	OH	M	Vasectomy	B1,B4
31	CA	F	Chemotherapy	B2,B5
35	OH	F	Hysterectomy	B2,B6

Attribute Ranking

- Step 1: compute attribute weights



Original table

Age	State	Gender	Treatment
…	…	…	…

Attribute Weights

Age	5
Gender	5
Treatment	4
State	0

Data Normalization

- Step 2: translate the original data table into a normalized table
  - Numerical attributes

Age	State	Gender	Treatment
42	OH	M	Vasectomy
47	OH	F	Hysterectomy
51	VA	F	Chemotherapy
55	VA	M	Chemotherapy
62	OH	M	Chemotherapy
67	CA	F	Hysterectomy
30	OH	M	Vasectomy
31	CA	F	Chemotherapy
35	OH	F	Hysterectomy

Age
42
47
51
55
62
67
30
31
35

Age (normalized)
0.63
0.7
0.76
0.82
0.93
1
0.45
0.46
0.52

Data Normalization

- Categorical attributes

Age	State	Gender	Treatment
42	OH	M	Vasectomy
47	OH	F	Hysterectomy
51	VA	F	Chemotherapy
55	VA	M	Chemotherapy
62	OH	M	Chemotherapy
67	CA	F	Hysterectomy
30	OH	M	Vasectomy
31	CA	F	Chemotherapy
35	OH	F	Hysterectomy

State
OH
OH
VA
VA
OH
CA
OH
CA
OH

OH	VA	CA
1	0	0
1	0	0
0	1	0
0	1	0
1	0	0
0	0	1
1	0	0
0	0	1
1	0	0

Data Normalization

Normalized table

Age (Normalized)	OH	VA	CA	Female	Male	Hysterectomy	Vasectomy	Chemotherapy
0.63	1	0	0	0	1	0	1	0
0.7	1	0	0	1	0	1	0	0
0.76	0	1	0	1	0	0	0	1
0.82	0	1	0	0	1	0	0	1
0.93	1	0	0	0	1	0	0	1
1	0	0	1	1	0	1	0	0
0.45	1	0	0	0	1	0	1	0
0.46	0	0	1	1	0	0	0	1
0.52	1	0	0	1	0	1	0	0

Clustering

- Step 3: apply weighted k-means clustering

Attribute Weights

Age	5
Gender	5
Treatment	4
State	0

Clustered table, k=3

Age (Normalized)	OH	VA	CA	Female	Male	Hysterectomy	Vasectomy	Chemotherapy
0.63	1	0	0	0	1	0	1	0
0.7	1	0	0	1	0	1	0	0
0.76	0	1	0	1	0	0	0	1
0.82	0	1	0	0	1	0	0	1
0.93	1	0	0	0	1	0	0	1
1	0	0	1	1	0	1	0	0
0.45	1	0	0	0	1	0	1	0
0.46	0	0	1	1	0	0	0	1
0.52	1	0	0	1	0	1	0	0

Computing centroid

- Step 4: compute the centroid records

0.63	1	0	0	0	1	0	1	0
0.7	1	0	0	1	0	1	0	0
0.76	0	1	0	1	0	0	0	1

0.7	0.67	0.33	0	0.67	0.33	0.33	0.33	0.33

Computing centroid

- Step 4: compute the centroid records

0.63	1	0	0	0	1	0	1	0
0.7	1	0	0	1	0	1	0	0
0.76	0	1	0	1	0	0	0	1

0.7	0.67	0.33	0	0.67	0.33	0.33	0.33	0.33

Centroid table

Age (Normalized)	OH	VA	CA	Female	Male	Hysterectomy	Vasectomy	Chemotherapy
0.7	0.67	0.33	0	0.67	0.33	0.33	0.33	0.33
0.92	0.33	0.33	0.33	0.33	0.67	0	0.33	0.67
0.48	0.67	0	0.33	0.67	0.33	0.33	0.33	0.33

Anonymized table

- Step 5: generate real anonymized table

Age (Normalized)	OH	VA	CA	Female	Male	Hysterectomy	Vasectomy	Chemotherapy
0.7	0.67	0.33	0	0.67	0.33	0.33	0.33	0.33

Age	State	Gender	Treatment
47	OH	F	Vasectomy

Anonymized table

Age	State	Gender	Treatment
47	OH	F	Vasectomy
61	VA	M	Chemotherapy
32	OH	F	Hysterectomy

Associative rule

- Step 6: generate and apply associative rule

Original table

Age	State	Gender	Treatment
42	OH	M	Vasectomy
47	OH	F	Hysterectomy
51	VA	F	Chemotherapy
55	VA	M	Chemotherapy
62	OH	M	Chemotherapy
67	CA	F	Hysterectomy
30	OH	M	Vasectomy
31	CA	F	Chemotherapy
35	OH	F	Hysterectomy

"Vasectomy->male"

Anonymized table

Age	State	Gender	Treatment
47	OH	F	Vasectomy
61	VA	M	Chemotherapy
32	OH	F	Hysterectomy

Anonymized table with correction

Age	State	Gender	Treatment
47	OH	M	Vasectomy
61	VA	M	Chemotherapy
32	OH	F	Hysterectomy

Branch Coverage

Original table

Age	State	Gender	Treatment	Branch
42	OH	M	Vasectomy	B1,B4
47	OH	F	Hysterectomy	B2,B6
51	VA	F	Chemotherapy	B2,B5
55	VA	M	Chemotherapy	B1,B3
62	OH	M	Chemotherapy	B1,B3
67	CA	F	Hysterectomy	B2,B6
30	OH	M	Vasectomy	B1,B4
31	CA	F	Chemotherapy	B2,B5
35	OH	F	Hysterectomy	B2,B6

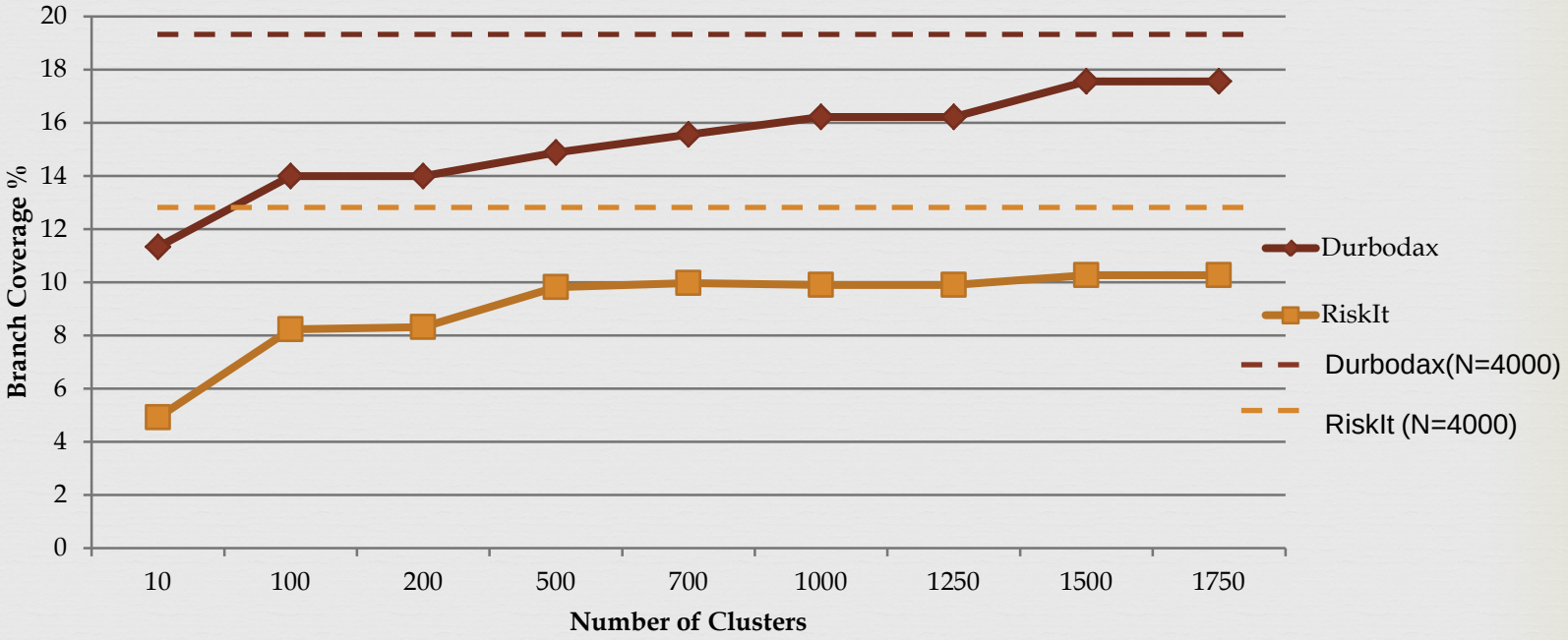
Anonymized table with correction

Age	State	Gender	Treatment	Branch
47	OH	M	Vasectomy	B1, B4
61	VA	M	Chemotherapy	B1, B3
32	OH	F	Hysterectomy	B2, B6

Experiment

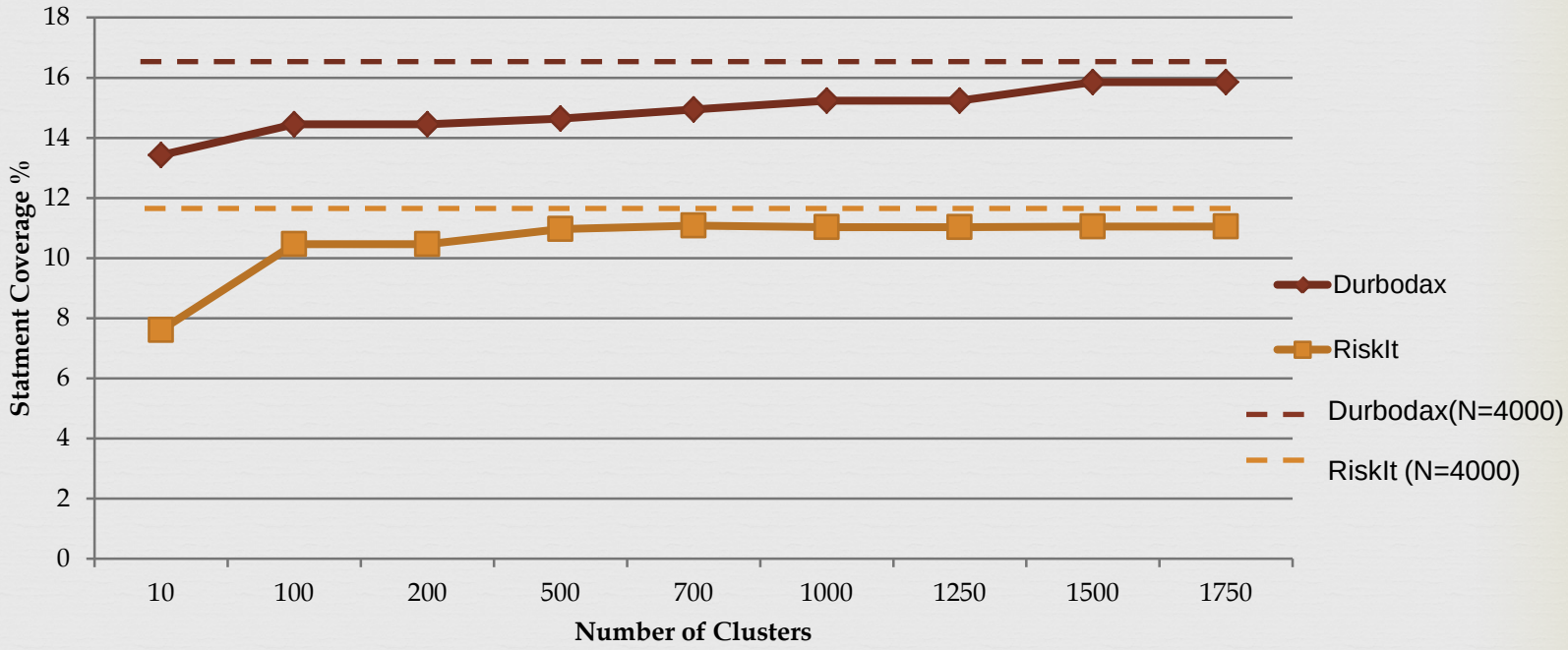
- We evaluated PISTIS on two open-source Java applications, DurboDax and RiskIt
  - DurboDax: 27 tables and 114 attributes
  - RiskIt: 14 tables and 57 attributes
  - Randomly select 4000 records
  - Branch coverage are 19.3% and 13% respectively.

Branch coverage



Branch coverage on the number of clusters for subject applications

Statement coverage



Statement coverage on the number of clusters for subject applications

Disclosure rate

- To evaluate privacy level
- Similarity matrix
- Disclosure rate: the average of all cells in the similarity matrix

Similarity matrix

Age	State	Gender	Treatment
42	OH	M	Vasectomy
47	OH	F	Hysterectomy
51	VA	F	Chemotherapy
55	VA	M	Chemotherapy
62	OH	M	Chemotherapy
67	CA	F	Hysterectomy
30	OH	M	Vasectomy
31	CA	F	Chemotherapy
35	OH	F	Hysterectomy

Original table

Age	State	Gender	Treatment
47	OH	M	Vasectomy
61	VA	M	Chemotherapy
32	OH	F	Hysterectomy

Anonymized table

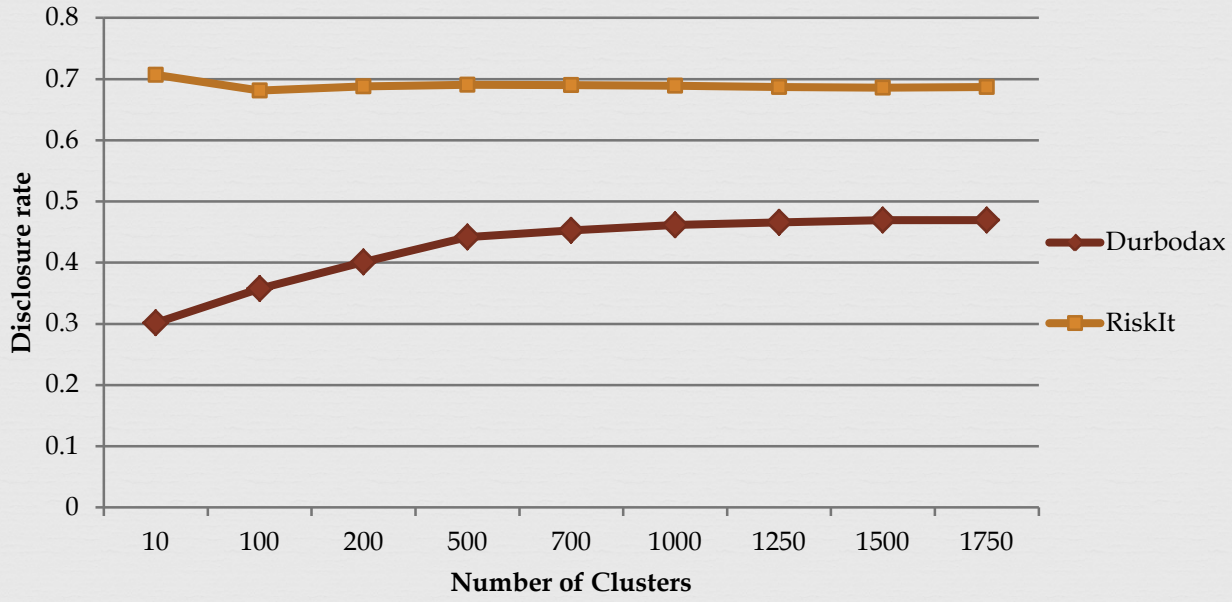
$$\frac{0 + 1 + 1 + 1}{4} = 0.75$$

- Similarity matrix

	Record 1	Record 2	Record 3	Record 4	Record 5	Record 6	Record 7	Record 8	Record 9
C1	0.75	0.25	0	0.25	0.5	0	0	0	0.25
C2	0.25	0	0.5	0.75	0.5	0	0.25	0.25	0
C3	0.25	0.75	0.25	0	0.25	0.5	0.25	0.25	0.75

Disclosure rate

- We compute the disclosure rate as the average of all cells in the similarity matrix



Disclosure rate on the number of clusters for subject applications

Summary

Our work

- We focus on balancing the following four dimensions:

Privacy	**Data minimization**
Semantic correctness	**Testing coverage**

6

PISTIS - Protecting and mInimizing databases for Software TestIng taSks

- Program analysis ------- **Test coverage**
- Weight-based k-clustering algorithm — **Data minimization**
- Compute centroid objects — **Privacy**
- Associative rule mining ------- **Semantic correctness**

PISTIS allows us to meet all four goals

12

Running example

- Step 4: maps the table back to the original attributes

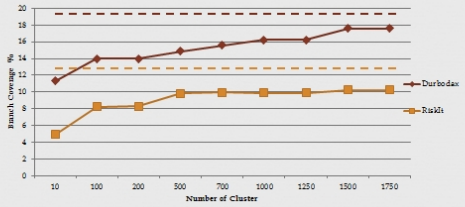
Age (Normalized)	OH	VA	CA	Female	Male	Hysterectomy	Vasectomy	Chemotherapy
0.7	0.67	0.33	0	0.67	0.33	0.33	0.33	0.33

Age	State	Gender	Treatment
47	OH	F	Vasectomy

Age	State	Gender	Treatment
47	OH	F	Vasectomy
61	VA	M	Chemotherapy
32	OH	F	Hysterectomy

18

Branch coverage



Branch coverage on the number of clusters for subject applications

29