Incognito: Efficient Full-Domain K-Anonymity

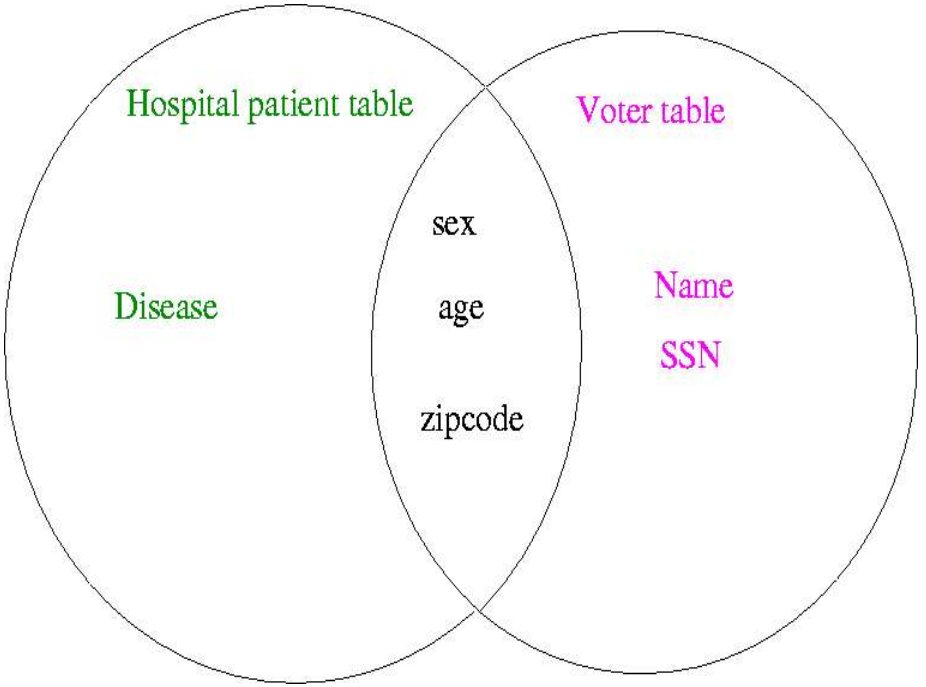
Kristen LeFevre David J. DeWitt Raghu Ramakrishnan University of Wisconsin Madison 1210 West Dayton St. Madison, WI 53706

Talk Prepared By
Parul Halwe(05305002)
Vibhooti Verma(05305016)

Motivation

- A number of organizations publish micro data for purposes such as public health and demographic research.
- It might lead to violation of data privacy of some individual.
- Some attribute that clearly identify individuals, such as Name and Social Security Number, are generally removed.

Just removing name and ssn are sufficient for data privacy?



- NO
- Databases can sometimes be joined with other public databases on attributes such as Zipcode, Sex, and Birthdate to re-identify individuals who were supposed to remain anonymous.

VOTER REGISTRATION DATA

Name	Age	Sex	Zipcode
Ahmed	25	Male	53711
Brooke	28	Female	55410
Claire	31	Female	90210
Dave	19	Male	02174
Evelyn	40	Female	02237

PATIENT DATA

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
25	Female	53712	Hepatitis
26	Male	53711	Brochitis
27	Male	53710	Broken Arm
27	Female	53712	AIDS
28	Male	53711	Hang Nail

Generalized Hospital table

VOTER REGISTRATION DATA

Name	Age	Sex	Zipcode
Ahmed	25	Male	53711
Brooke	28	Female	55410
Claire	31	Female	90210
Dave	19	Male	02174
Evelyn	40	Female	02237

PATIENT DATA

Age	Sex	Zipcode	Disease
2*	Male	53711	Flu
2*	Female	53712	Hepatitis
2*	Male	53711	Brochitis
2*	Male	53710	Broken Arm
2*	Female	53712	AIDS
2*	Male	53711	Hang Nail

How can we make individual's data private along with publishing Microdata?

- K-Anonymity : K-anonymization is a technique that prevents joining attacks by generalizing and/or suppressing portions of the released microdata so that no individual can be uniquely distinguished from a group of size k.

Example of generalized table for k=2

VOTER REGISTRATION DATA

Name	Age	Sex	Zipcode
Ahmed	25	Male	53711
Brooke	28	Female	55410
Claire	31	Female	90210
Dave	19	Male	02174
Evelyn	40	Female	02237

PATIENT DATA

Age	Sex	Zipcode	Disease
2*	Male	5371*	Flu
2*	Female	5371*	Hepatitis
2*	Male	5371*	Brochitis
2*	Male	5371*	Broken Arm
2*	Female	5371*	AIDS
2*	Male	5371*	Hang Nail

- Generalize age and zipcode by one digit

Terminologies

- Quasi-Identifier Attribute Set :A quasi-identifier set Q is a minimal set of attributes in table T that can be joined with external information to re-identify individual records.
- Frequency Set :. The frequency set of T with respect to Q is a mapping from each unique combination of values (q0..... qn) of Q in T (the value groups) to the total number of tuples in T with these values of Q (the counts).
- K-Anonymity Property: Relation T is said to satisfy the k-anonymity property (or to be k-anonymous) with respect to attribute set Q if every count in the frequency set of T with respect to Q is greater than or equal to k.

K-anonymization Techniques

- Generalization : Generalization of domain values of relational attributes to more general values.
- Suppression : Dropping some tuples from relation to satisfy k-anonymity

No Generalization

Table B0

PATIENT DATA

Birthday	Sex	Zipcode	Disease
1/21/76	Male	53711	Flu
4/13/86	Female	53711*	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	AIDS
1/28/76	Male	53706	Hang Nail

Generalization on Birthday

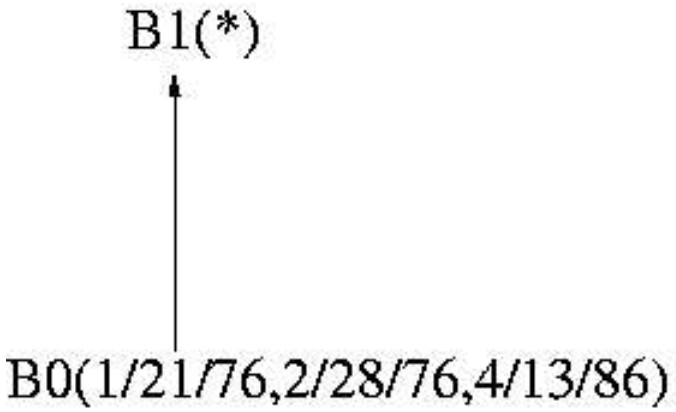
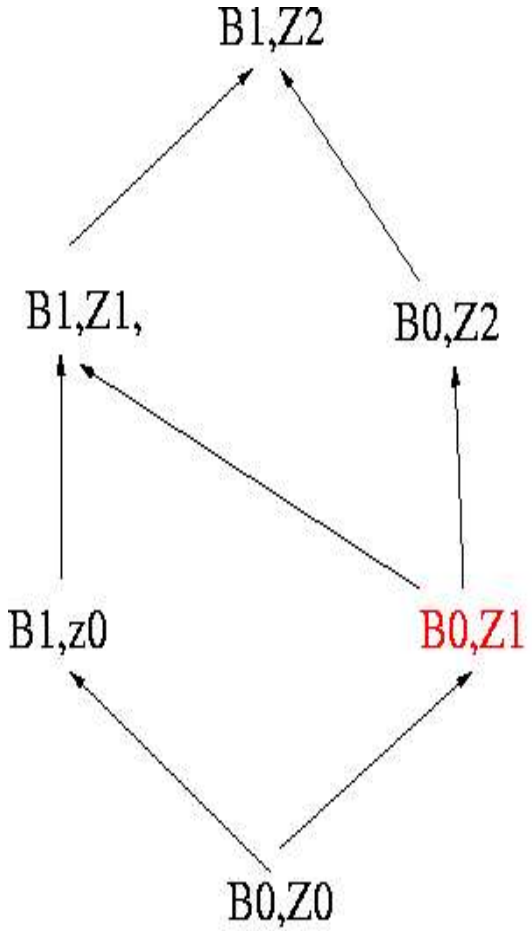


Table B1

PATIENT DATA

Birthday	Sex	Zipcode	Disease
*	Male	53715	Flu
*	Female	53715	Hepatitis
*	Male	53703	Brochitis
*	Male	53703	Broken Arm
*	Female	53706	AIDS
*	Male	53706	Hang Nail

Generalization on Birthday and Zipcode

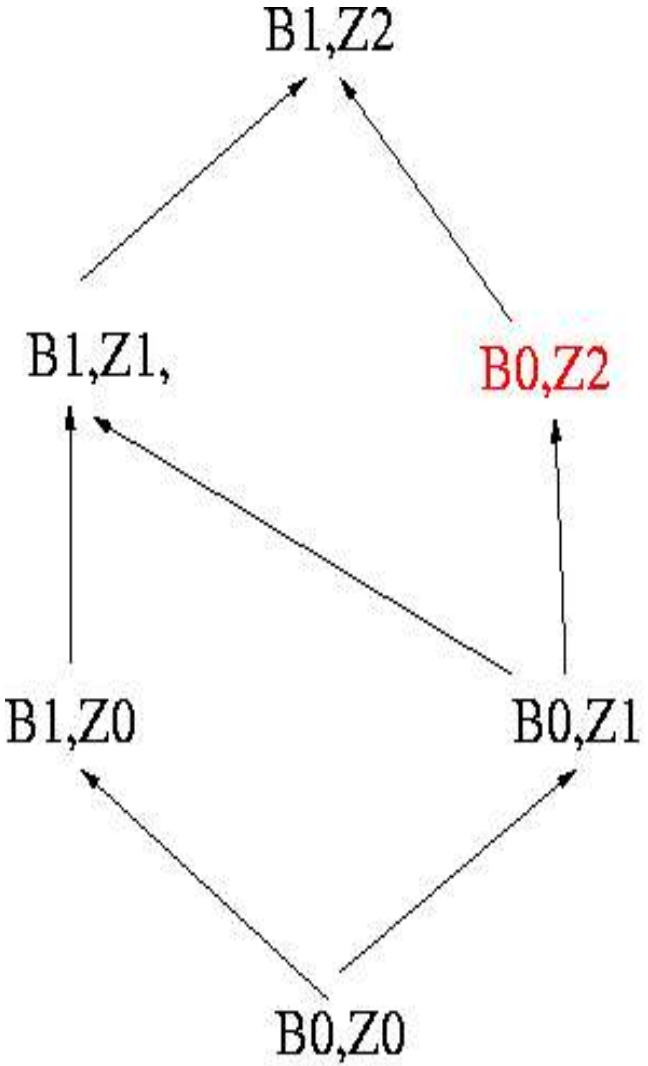


PATIENT DATA

Birthday	Sex	Zipcode	Disease
1/21/76	Male	5371*	Flu
4/13/86	Female	5371*	Hepatitis
2/28/76	Male	5370*	Brochitis
1/21/76	Male	5370*	Broken Arm
4/13/86	Female	5370*	AIDS
1/28/76	Male	5370*	Hang Nail

Table B0,Z1

Generalization on Birthday and Zipcode

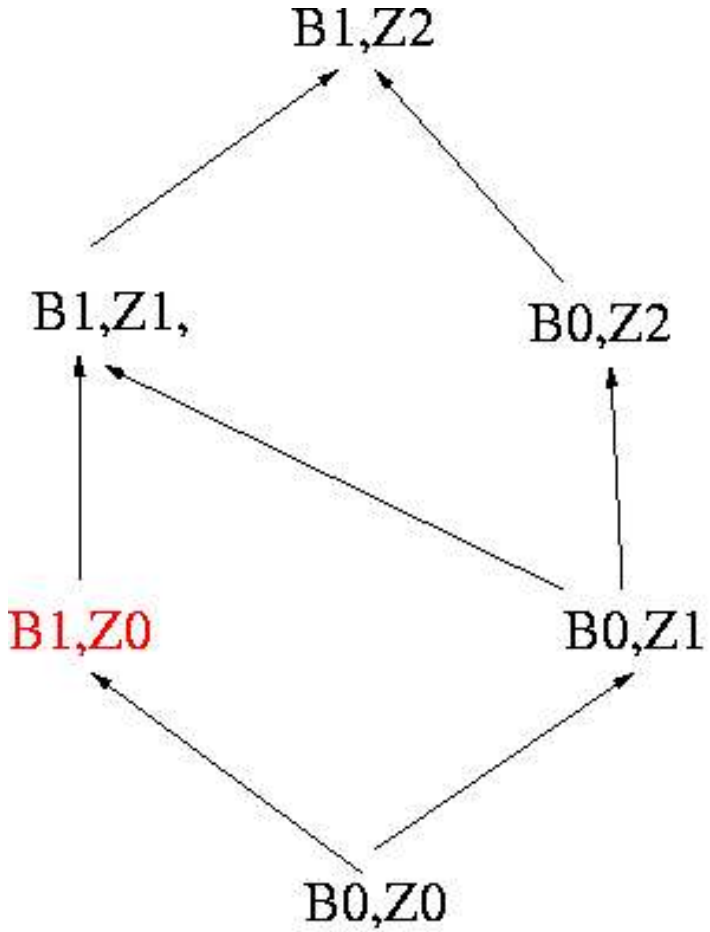


PATIENT DATA

Birthday	Sex	Zipcode	Disease
1/21/76	Male	537**	Flu
4/13/86	Female	537**	Hepatitis
2/28/76	Male	537**	Brochitis
1/21/76	Male	537**	Broken Arm
4/13/86	Female	537**	AIDS
1/28/76	Male	537**	Hang Nail

Table B0,Z2

Generalization on Birthday and

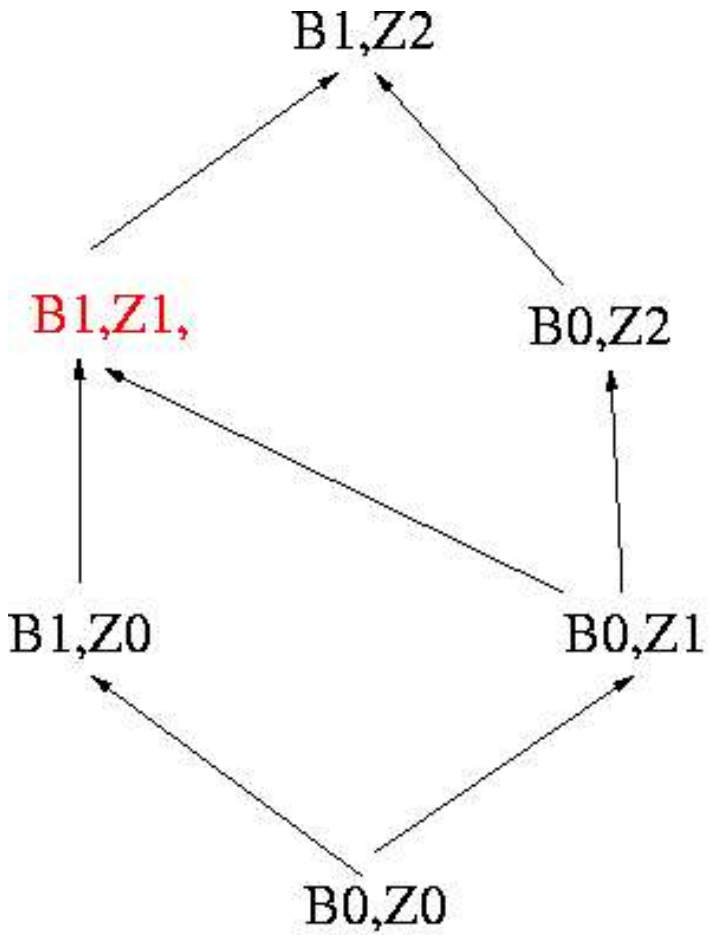


PATIENT DATA

Birthday	Sex	Zipcode	Disease
*	Male	53711	Flu
*	Female	53711	Hepatitis
*	Male	53703	Brochitis
*	Male	53703	Broken Arm
*	Female	53706	AIDS
*	Male	53706	Hang Nail

Table B1,Z0

Generalization on Birthday and Zipcode

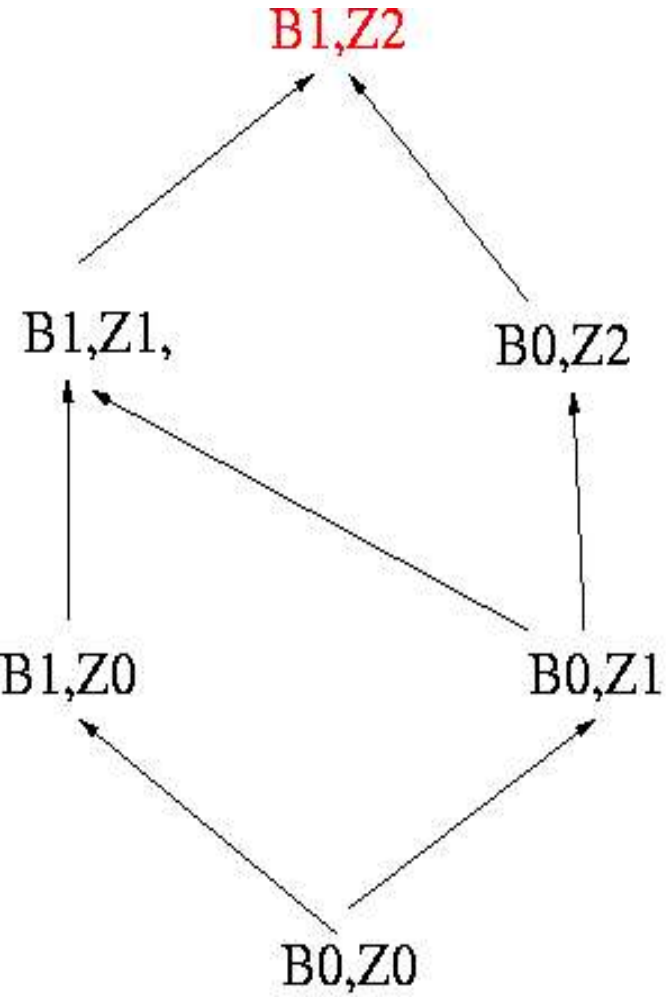


PATIENT DATA

Birthday	Sex	Zipcode	Disease
*	Male	5371*	Flu
*	Female	5371*	Hepatitis
*	Male	5370*	Brochitis
*	Male	5370*	Broken Arm
*	Female	5370*	AIDS
*	Male	5370*	Hang Nail

Table B1,Z1

GENERALIZATION



PATIENT DATA

Birthday	Sex	Zipcode	Disease
*	Male	537**	Flu
*	Female	537**	Hepatitis
*	Male	537**	Brochitis
*	Male	537**	Broken Arm
*	Female	537**	AIDS
*	Male	537**	Hang Nail

Table B1,Z2

Domain Generalization

- Domain Generalization Relationship : Let Ti(a1...an) and Tj(a1....an) be 2 tables defined on same set of attributes.Then Tj will be called generalization of Ti(Ti <d Tj) iff
- |Ti| = |Tj|
- For all z for z=1.....n , dom(A_z ,T_j) <= (A_z T_i)
- It is possible to define a bijective mapping between Ti and Tj that associate each tuple ti and tj such that $t_j[A_z] <= t_i[A_z]$.

Genralization for Hospital patient Data

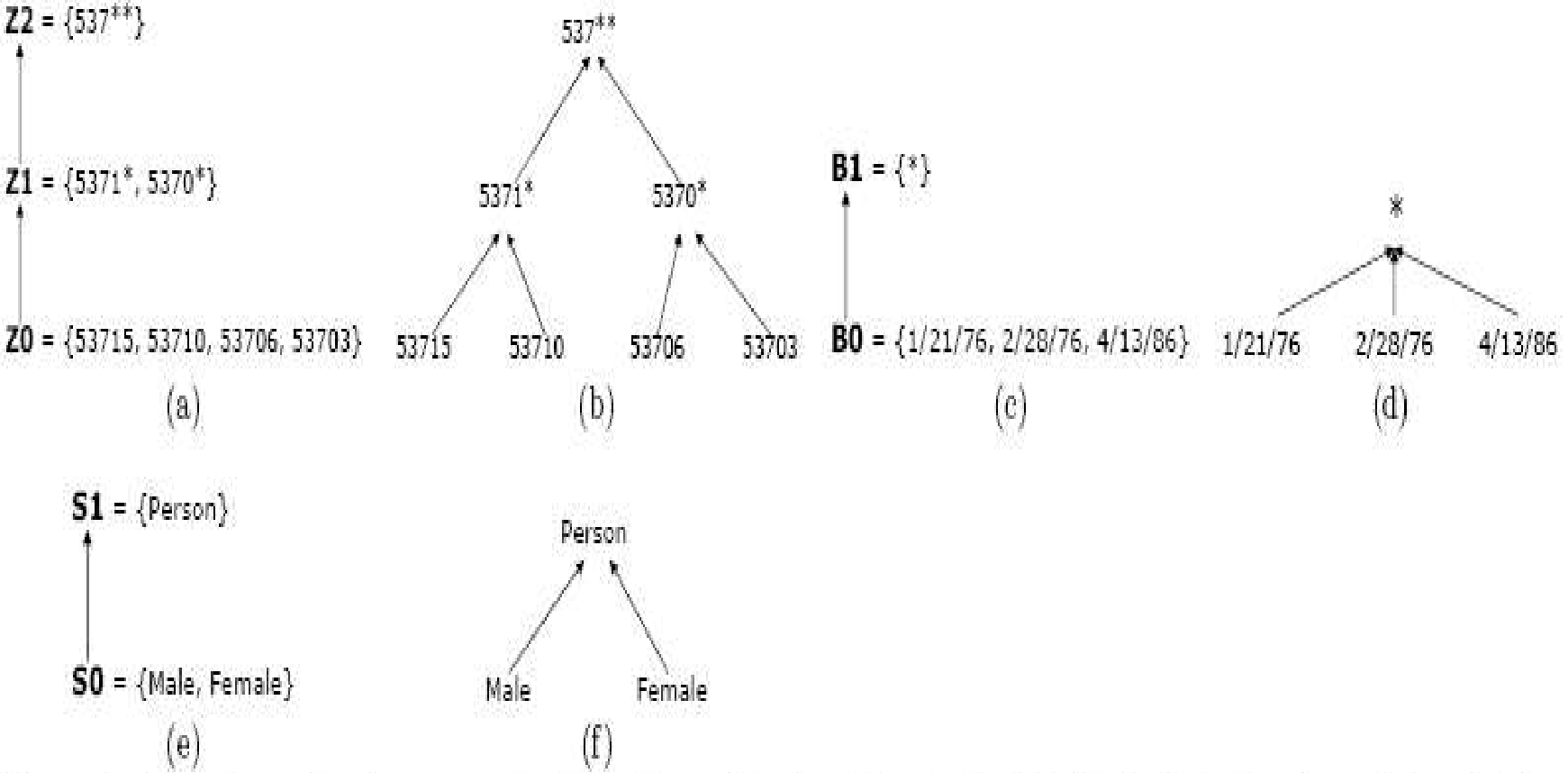
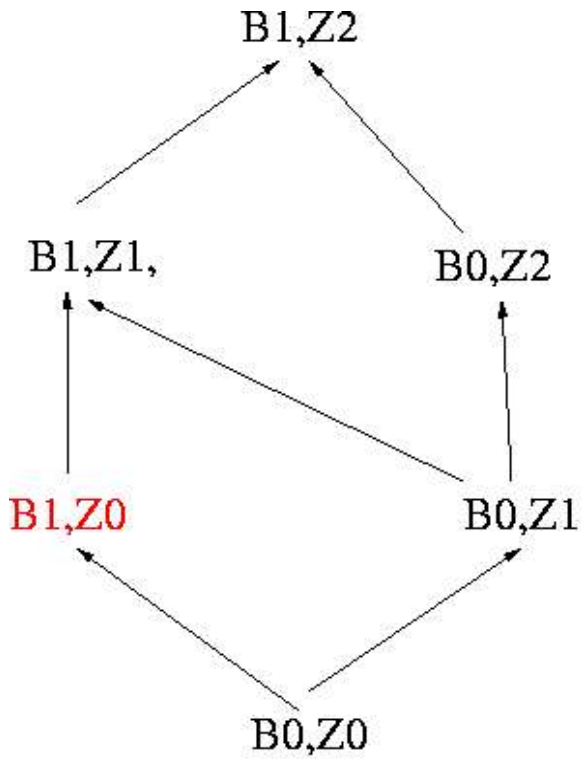


Figure 2: Domain and value generalization hierarchies for Zipcode (a, b), Birth Date (c, d), and Sex (e, f)

Suppression



PATIENT DATA

Birthday	Sex	Zipcode	Disease
*	Male	53711	Flu
*	Female	53711	Hepatitis
*	Male	53703	Brochitis
*	Male	53703	Broken Arm
*	Female	53706	AIDS
*	Male	53706	Hang Nail
*	Male	23567	Flu

Table B1,Z0

- Removing data from the table so that they are not released

Generalization for achieving 2 anonymity

PATIENT DATA

Birthday	Sex	Zipcode	Disease
*	Male	5*	Flu
*	Female	5*	Hepatitis
*	Male	5*	Brochitis
*	Male	5*	Broken Arm
*	Female	5*	AIDS
*	Male	5*	Hang Nail
*	Male	2*	Flu

Table B1,Z4

Suppressing 1 tuple to achieve 2-anonymity

PATIENT DATA

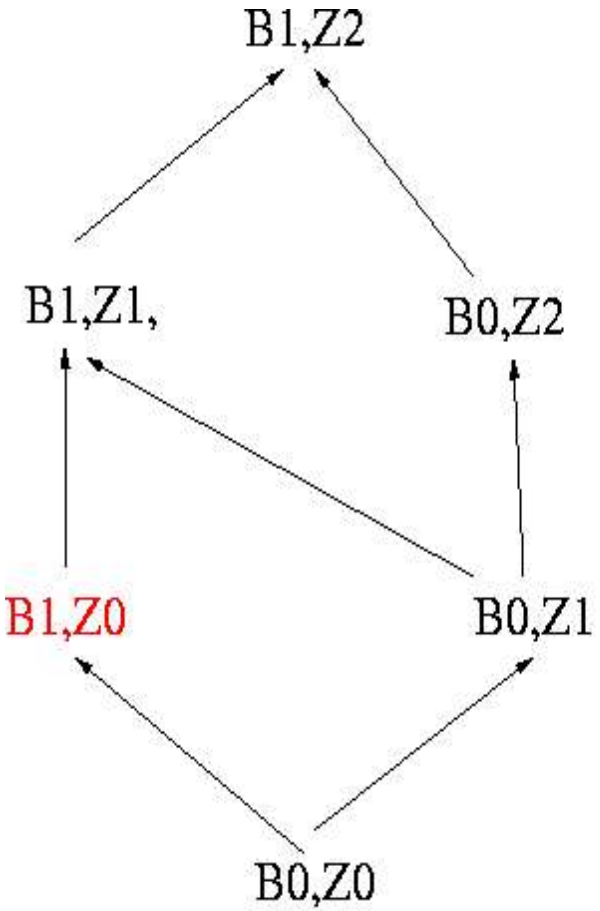
Birthday	Sex	Zipcode	Disease
*	Male	5*	Flu
*	Female	5*	Hepatitis
*	Male	5*	Brochitis
*	Male	5*	Broken Arm
*	Female	5*	AIDS
*	Male	5*	Hang Nail
*	Male	2*	Flu

Table B1,Z4

K-Minimal Generalization

- K-Minimal Generalization: let Ti and Tj be two tables such that Ti<Tj . Tj will said to be k-minimal generalization of Ti iff
  1. Tj satisfies k-anonymity
  2.There exist no Tz such that Ti<Tz , Tz satisfies k-anonymity and Di,j < Di,z

GENERALIZATION ON BIRTHDAY AND ZIPCODE FOR K=2(minimal)

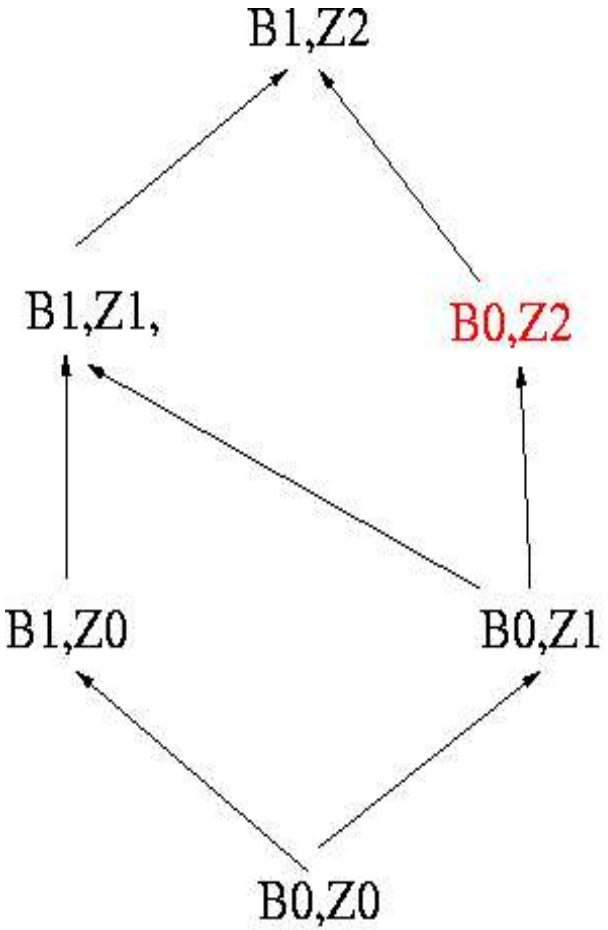


PATIENT DATA

Birthday	Sex	Zipcode	Disease
*	Male	53711	Flu
*	Female	53711	Hepatitis
*	Male	53703	Brochitis
*	Male	53703	Broken Arm
*	Female	53706	AIDS
*	Male	53706	Hang Nail

Table B1,Z0

GENERALIZATION ON BIRTHDAY AND ZIPCODE FOR K=2(not minimal)



PATIENT DATA

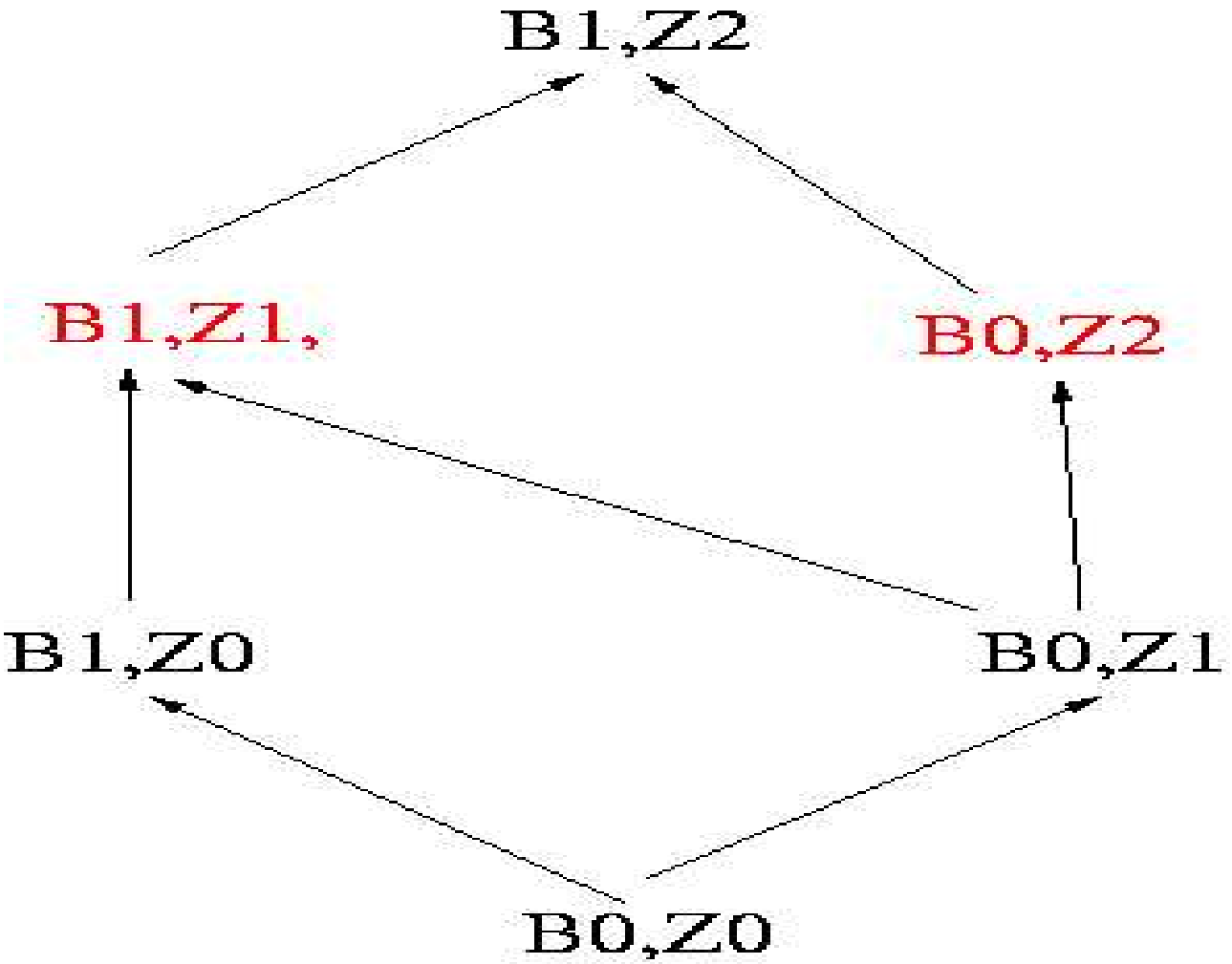
Birthday	Sex	Zipcode	Disease
1/21/76	Male	537**	Flu
4/13/86	Female	537**	Hepatitis
2/28/76	Male	537**	Brochitis
1/21/76	Male	537**	Broken Arm
4/13/86	Female	537**	AIDS
1/28/76	Male	537**	Hang Nail

Table B0,Z2

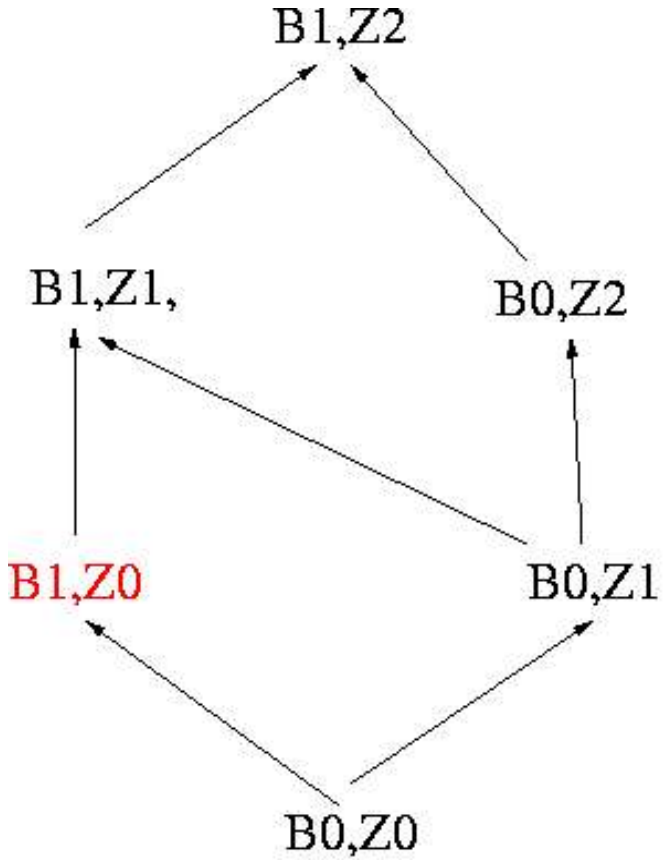
Full Domain Generalization Algorithms

- Binary Search
- Bottom up without Rollup
- Bottom up with Rollup
- Basic Incognito
- Super-roots Incognito

Binary Search



Bottom up with Roll- up



PATIENT DATA

Birthday	Sex	Zipcode	Disease
*	Male	53711	Flu
*	Female	53711	Hepatitis
*	Male	53703	Brochitis
*	Male	53703	Broken Arm
*	Female	53706	AIDS
*	Male	53706	Hang Nail

Table B1,Z0

Full Domain Generalization Properties

- Generalization Property :Let T be a relation, and P and Q be sets of attributes in T such that $D_P < D_Q$. If T is k-anonymous with respect to P, then T is also anonymous with respect to Q.
- Rollup Property : Let T be a relation, and let P and Q be sets of attributes such that DP <= DQ. If we have f1, the frequency set of T with respect to P, then we can generate each count in f2, the frequency set of T with respect to Q, by summing the set of counts in f1 associated by r with each value set of f2 .
- Subset Property: Let T be a relation, and let Q be a set of attributes in T. If T is k-anonymous with respect to Q, then T is k-anonymous with respect to any set of attributes P such that P <= Q.

Basic Incognito Algorithm

- Each iteration considers a graph of candidate multi- attribute generalization (nodes) constructed from a subset of the quasi-identifier of size i.
- A modified breadth first search over the graph yields the set of multi-attribute generalization of size i with respect to which T is K anonymous.
- After obtaining Si, the algorithm constructs the set of candidate nodes of size i + 1 (Ci+1), and the edges connecting them (Ei+1) using the subset property.

Graph Construction

1. Join Phase : It creates a superset of Ci based on Si-1.
2. Prune Phase : a prune phase for generating the set of candidate nodes Ci with respect to which T could potentially be k-anonymous given previous iterations.
3. Edge Generation :Through this direct multi-attribute generalization relationships among candidate nodes are constructed.

Step 1

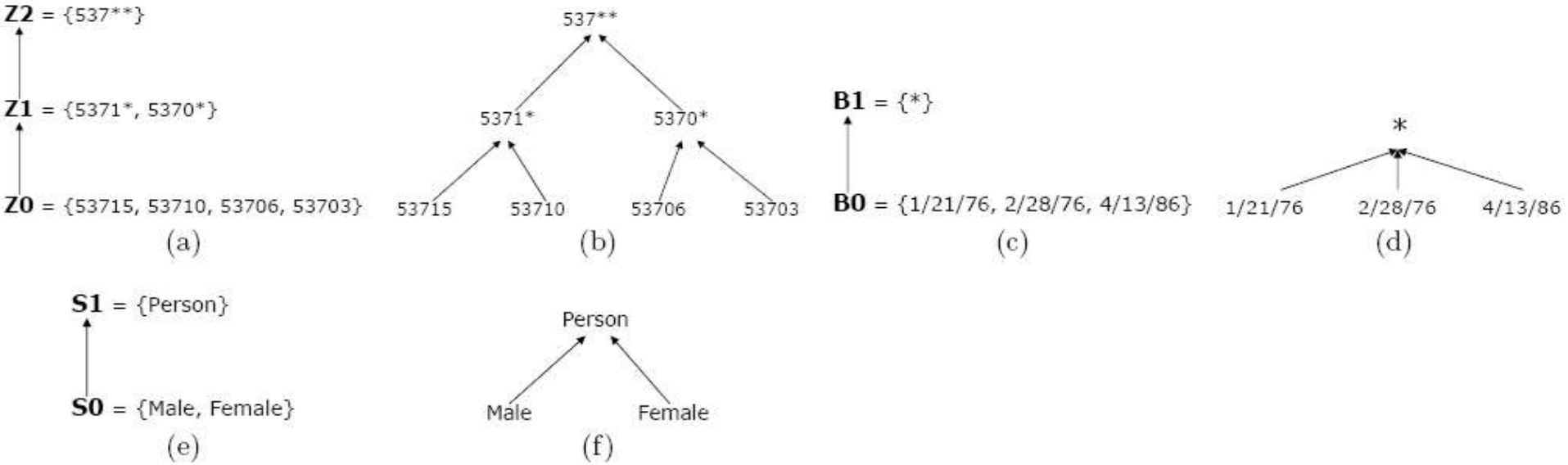


Figure 2: Domain and value generalization hierarchies for Zipcode (a, b), Birth Date (c, d), and Sex (e, f)

Step 2

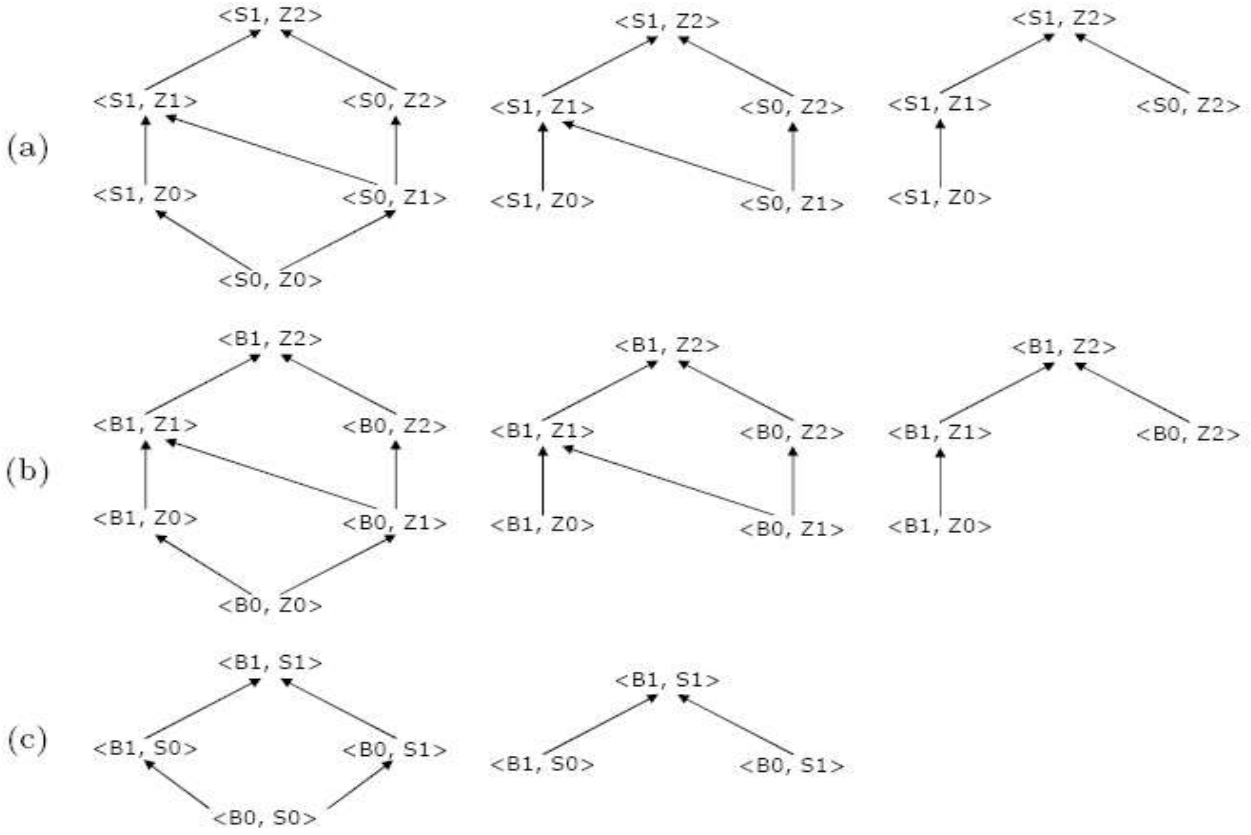
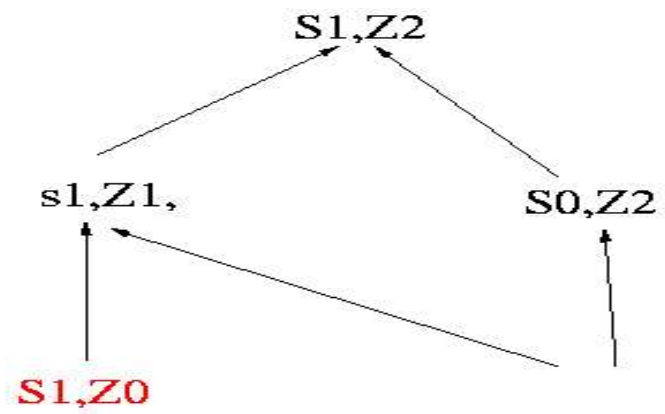
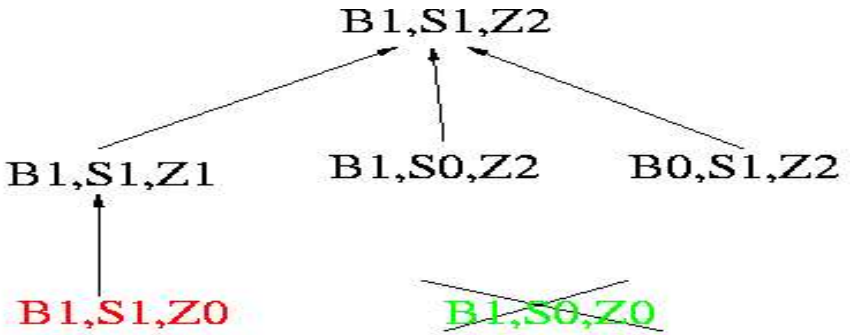
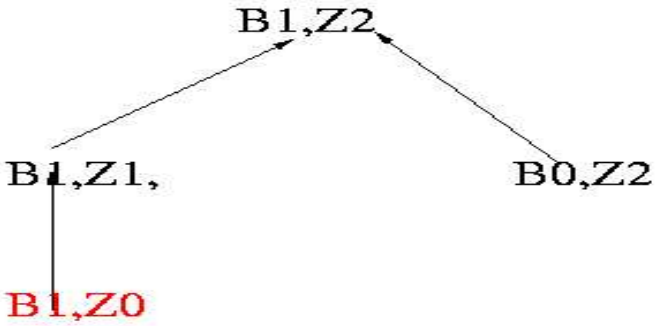


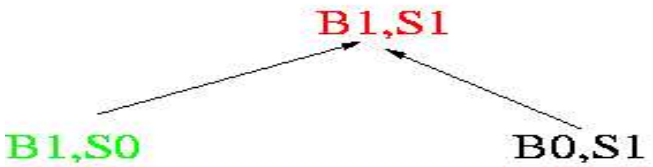
Figure 5: Searching the candidate 2-attribute generalization graphs for Patients example (Figure 1)

Step 3

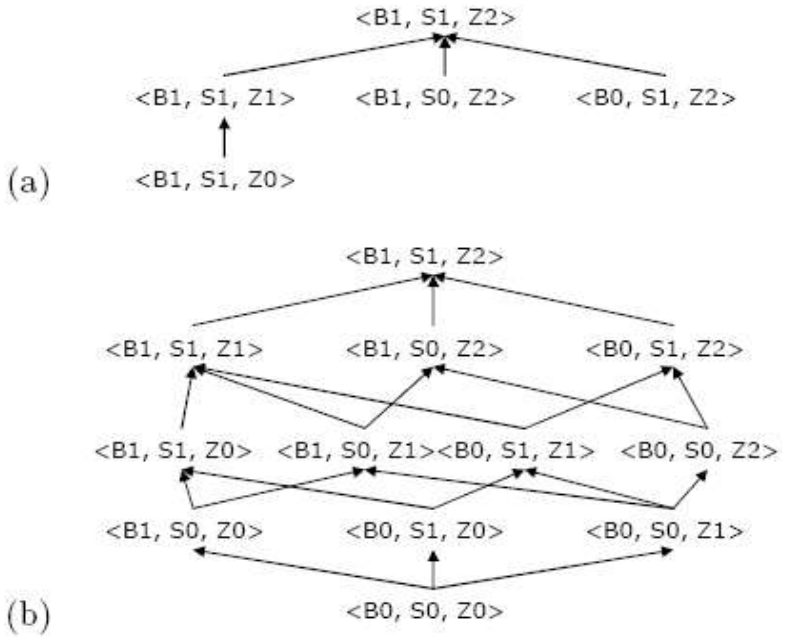

S1,Z2
s1,Z1,
S0,Z2
S1,Z0


B1,S1,Z2
B1,S1,Z1
B1,S0,Z2
B0,S1,Z2
B1,S1,Z0
B1,S0,Z0


B1,Z2
B1,Z1,
B0,Z2
B1,Z0

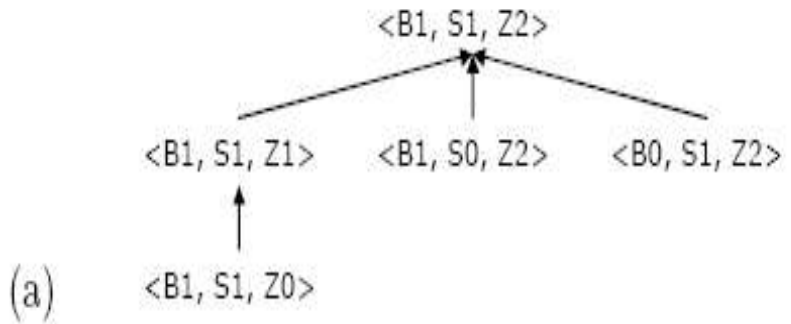

B1,S1
B1,S0
B0,S1

Comparison between Incognito and Bottom up algorithm



Algorithm Optimization

1. Super -roots : It is more efficient to group roots according to family, and then scan the database once,generating the frequency set corresponding to the least upper bound of each group (the \super-root").



Bottom up Pre-computation

- Here we generate the frequency sets of T with respect to all subsets of the quasi-identifier at the lowest level of generalization.
- Bottom up aggregation can be used.
- To overcome the fundamental drawback to of a priori optimizations , where single-attribute subsets are processes first.
- Example: we can not use the frequency set of T with respect to (Zipcode) to generate the frequency set of T with respect to (Sex, Zipcode).
- On the other hand, in the context of computing the data cube, these group-by queries would be processed in the opposite order, and rather than re-scanning the database, we could compute the frequency set of T with respect to (Zipcode) by simply rolling up the frequency set with respect to (Sex, Zipcode).

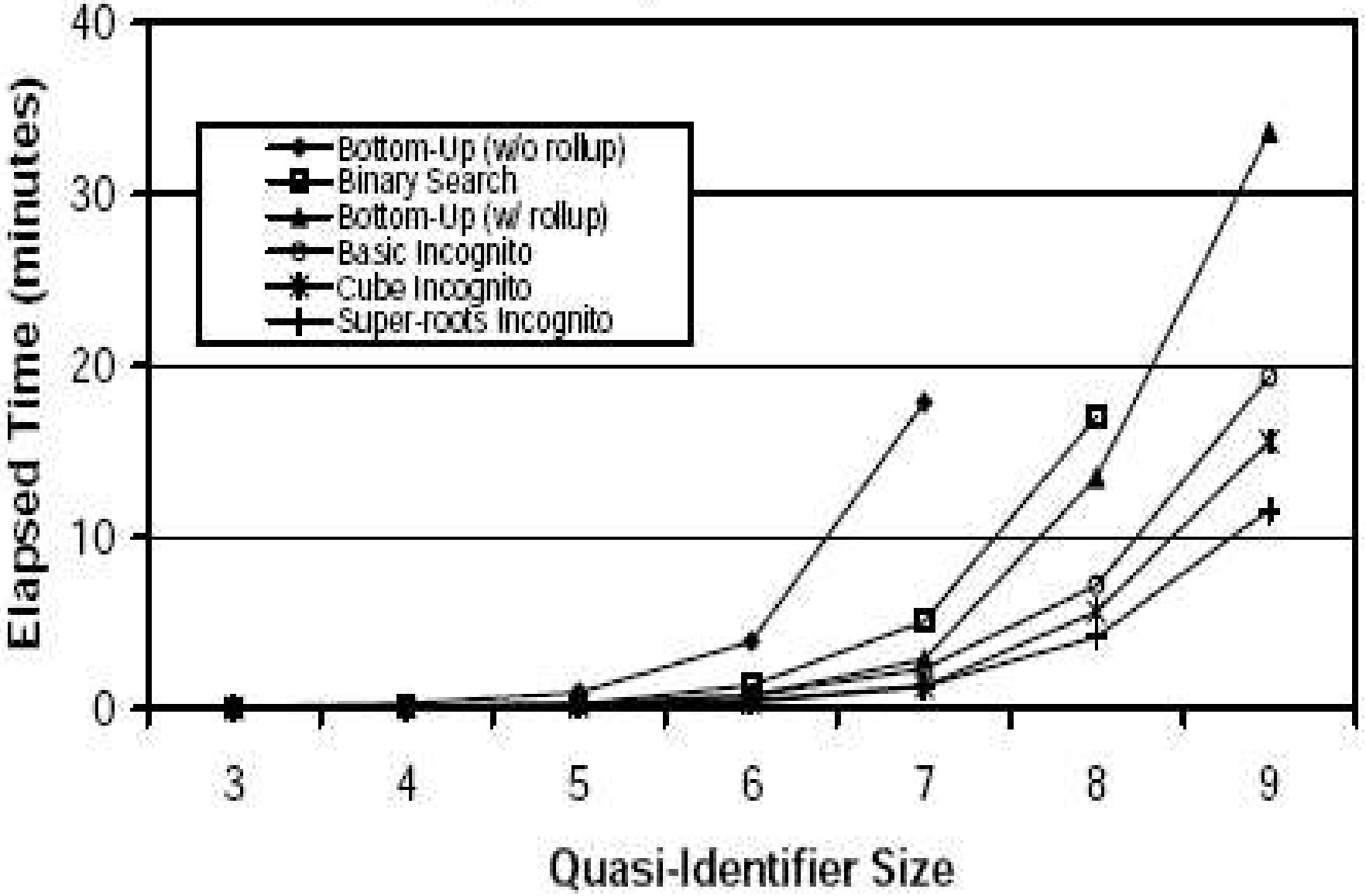
Experimental Data and Setup

- Adults database from the UC Irvine Machine Learning Repository which is comprised of data from the US Census.45000 records(5.5 MB)
- Lands End Corporation(4,591,581 records (268MB)
- AMD Athlon 1.5 GHz machine with 2 GB physical memory
- Microsoft windows 2003
- DB2 Enterprise Server Edition Version 8.1.2.
- The buffer pool size was set to 256 MB.

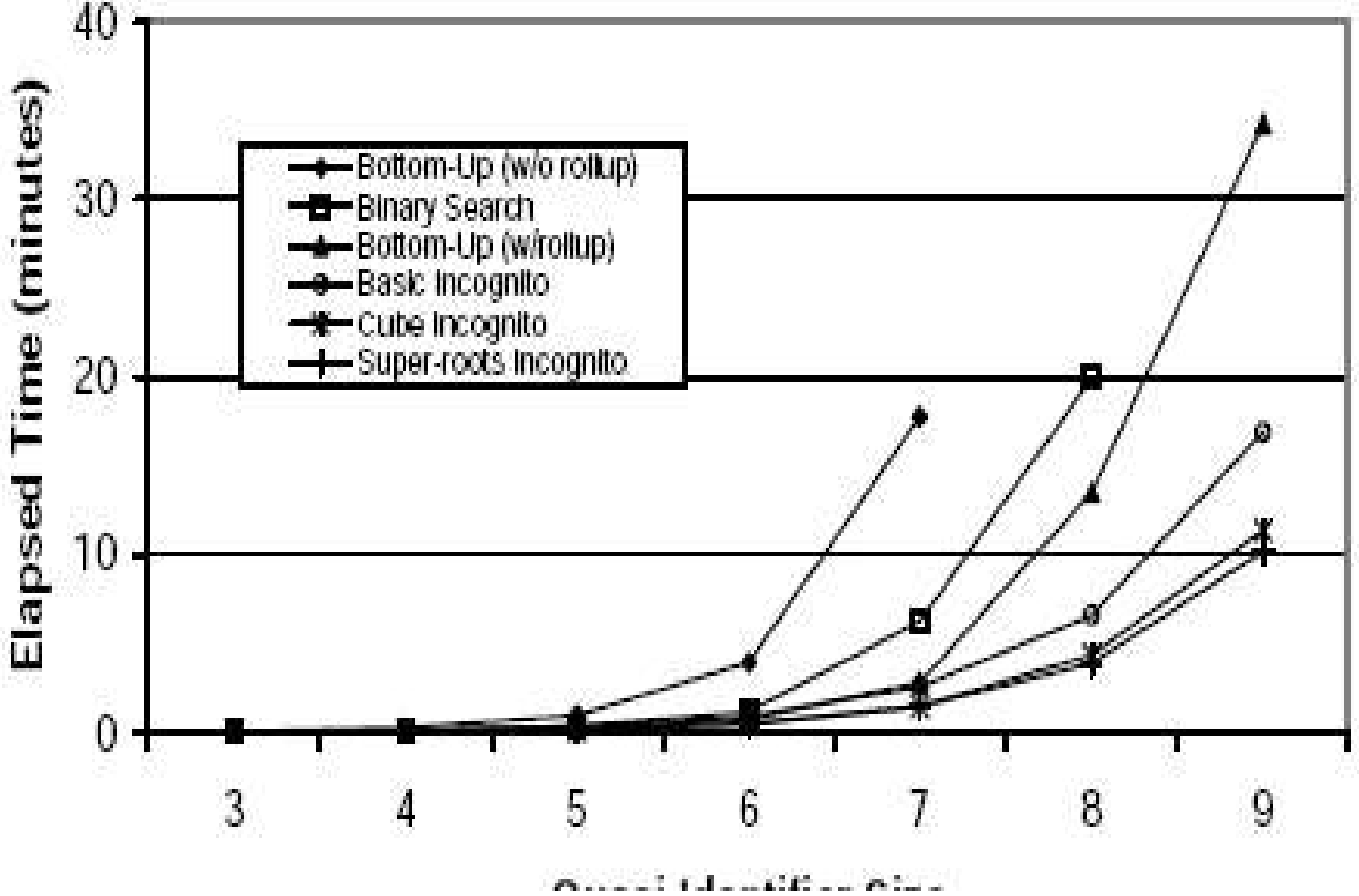
Experiment Results

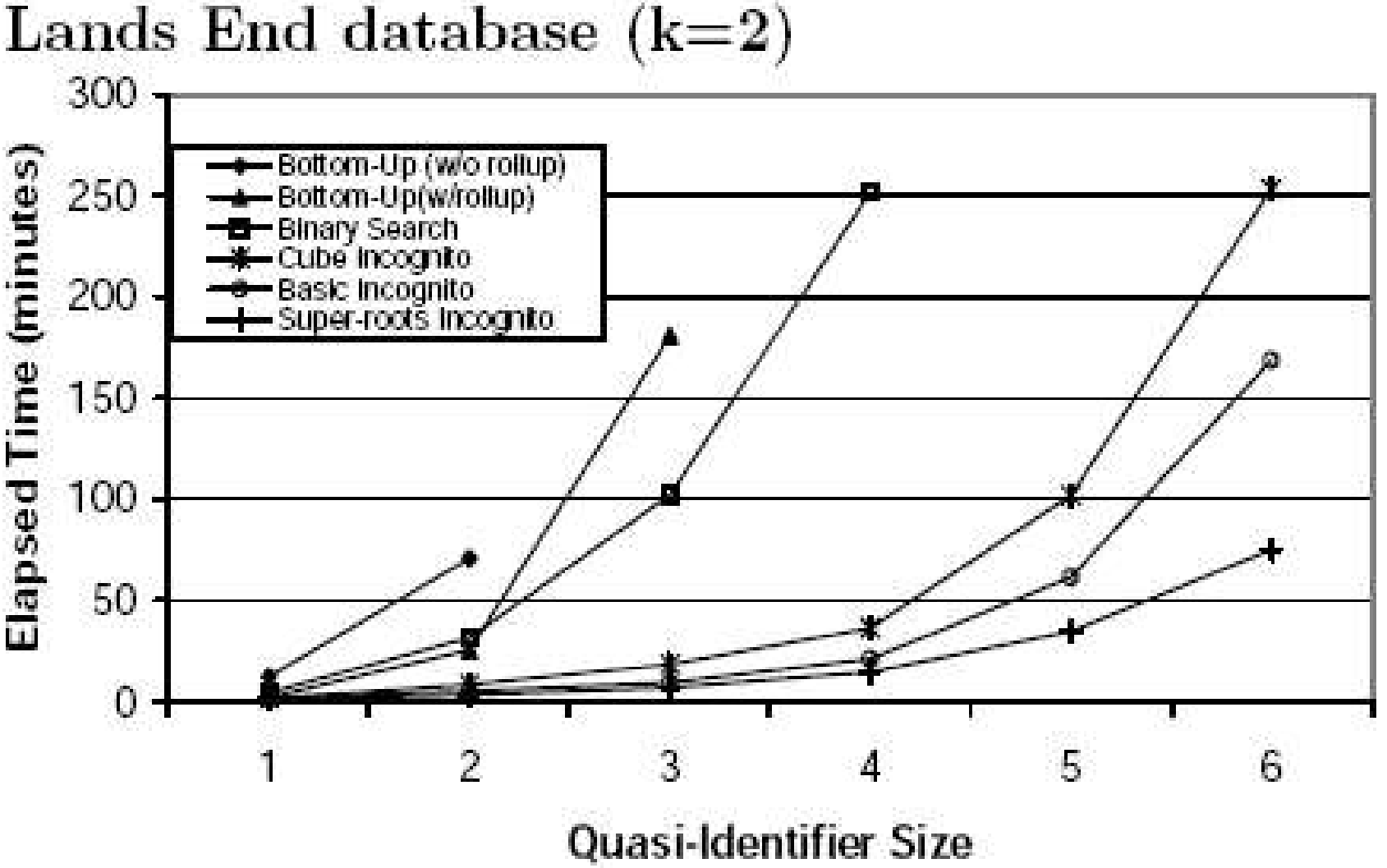
QID size	Bottom-Up	Incognito
3	14	14
4	47	35
5	206	103
6	680	246
7	2088	664
8	6366	1778
9	12818	4307

Adults database (k=2)

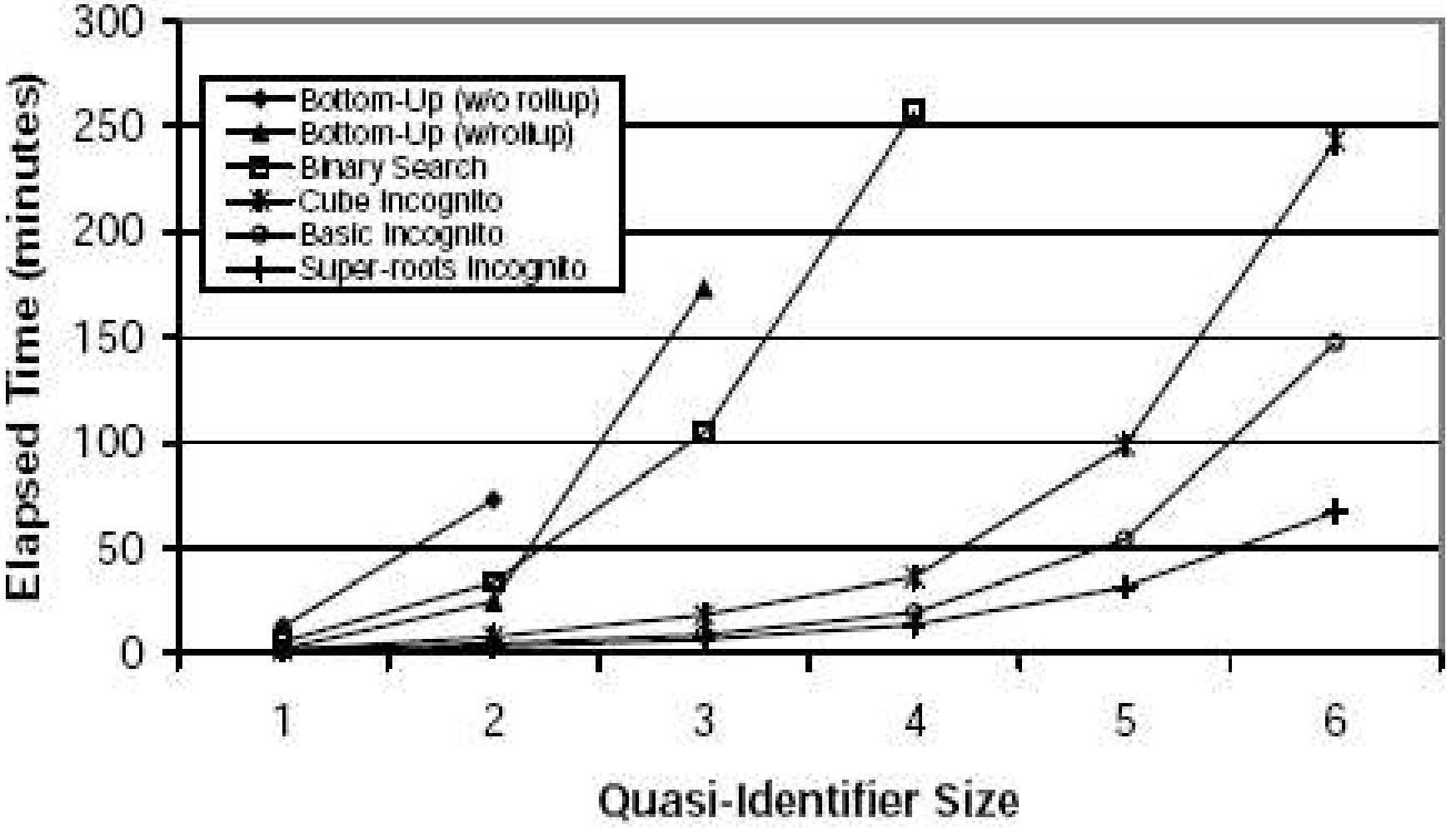


Adults database (k=10)




Lands End database (k=2)
Elapsed Time (minutes)
300
250
200
150
100
50
0
Bottom-Up (w/o rollup)
Bottom-Up(w/rollup)
Binary Search
Cube Incognito
Basic Incognito
Super-roots Incognito
1
2
3
4
5
6
Quasi-Identifier Size

Lands End database (k=10)



Mondrian Multidimensional K-Anonymity

Kristen Lefevre David J. DeWitt
Raghu Ramakrishna
University of Wisconsin, Madison

IEEE, ICDE 2006
(Previously UW Technical Report)

What could be another method for Anonymization?

- We can partition the domain into ranges rather than generalizing the values.
- This can be done for attributes which have a totally ordered domain.
- Each attribute can be viewed as a dimension.

Some Definitions

- Global Recoding
  - Single-dimensional Global Recoding: Defined by function

$$\Phi_i : D_{X_i} \rightarrow D'$$

  - Multidimensional Global Recoding: Defined by one function

$$\Phi : D_{X_1} \times ..... \times D_{X_n} \rightarrow D'$$

- Single-dimensional Partitioning
  - For each attribute define non-overlapping partitions for domain values
- Strict Multidimensional Partitioning
  - A multidimensional region is defined by pair of d-tuples

$$(p_1, ..., p_d)(v_1, ..., v_d) \in D_{X_i} \times ..... \times D_{X_d}$$

Contributions of this paper

- They propose a new multidimensional recoding model for k-anonymization and a greedy algorithm for this model.
- The greedy algorithm is more efficient than proposed algorithms for single-dimensional model.
- The greedy algorithm often produces higher-quality results than optimal single-dimensional algorithms.

An Example to Show Multidimensional Partitioning

VOTER REGISTRATION DATA

Name	Age	Sex	Zipcode
Ahmed	25	Male	53711
Brooke	28	Female	55410
Claire	31	Female	90210
Dave	19	Male	02174
Evelyn	40	Female	02237

PATIENT DATA

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
25	Female	53712	Hepatitis
26	Male	53711	Brochitis
27	Male	53710	Broken Arm
27	Female	53712	AIDS
28	Male	53711	Hang Nail

PATIENT DATA

Age	Sex	Zipcode	Disease
[25 – 27]	Male	53711	Flu
[25 – 27]	Female	53712	Hepatitis
[25 – 27]	Male	53711	Brochitis
[25 –27]	Male	53710	Broken Arm
[25 – 27]	Female	53712	AIDS
28	Male	53711	Hang Nail

Single-dimensional partitioning

PATIENT DATA

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
[25 – 27]	Female	53712	Hepatitis
26	Male	53711	Brochitis
27	Male	53710	Broken Arm
[25 – 27]	Female	53712	AIDS
28	Male	53711	Hang Nail

Multidimensional partitioning

PATIENT DATA

Age	Sex	Zipcode	Disease
[25 – 28]	Male	[53710 – 53711]	Flu
[25 – 28]	Female	53712	Hepatitis
[25 – 28]	Male	[53710 – 53711]	Brochitis
[25 – 28]	Male	[53710 – 53711]	Broken Arm
[25 – 28]	Female	53712	AIDS
[25 – 28]	Male	[53710 – 53711]	Hang Nail

Single-dimensional partitioning

PATIENT DATA

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
[25 – 27]	Female	53712	Hepatitis
26	Male	53711	Brochitis
[27 – 28]	Male	53710	Broken Arm
[25 – 27]	Female	53712	AIDS
[27 – 28]	Male	53711	Hang Nail

Multidimensional partitioning

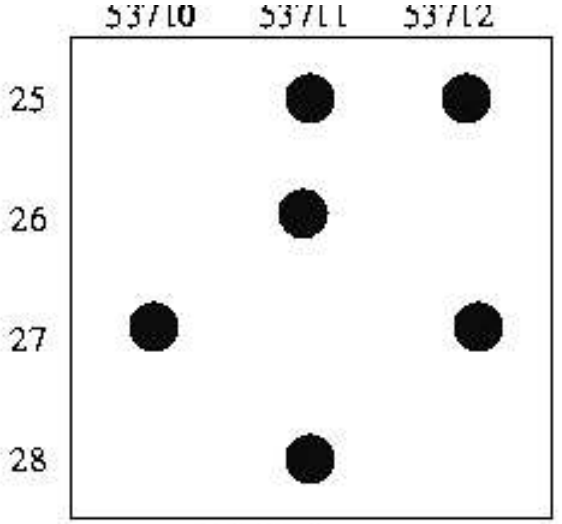
Age	Sex	Zipcode	Disease
[25 – 28]	Male	[53710 – 53711]	Flu
[25 – 28]	Female	53712	Hepatitis
[25 – 28]	Male	[53710 – 53711]	Brochitis
[25 – 28]	Male	[53710 – 53711]	Broken Arm
[25 – 28]	Female	53712	AIDS
[25 – 28]	Male	[53710 – 53711]	Hang Nail

Single-dimensional partitioning

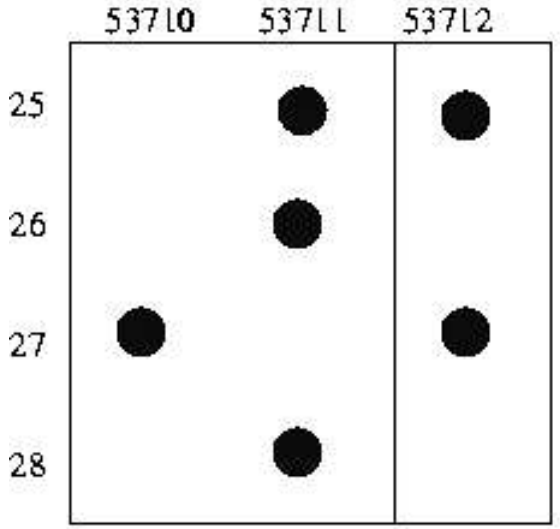
Age	Sex	Zip code	Disease
[25 – 26]	Male	53711	Flu
[25 – 27]	Female	53712	Hepatitis
[25 – 26]	Male	53711	Bronchitis
[27 – 28]	Male	[53710 – 53711]	Broken Arm
[25 – 27]	Female	53712	AIDS
[27 – 28]	Male	[53710 – 53711]	Hang Nail

Multidimensional partitioning

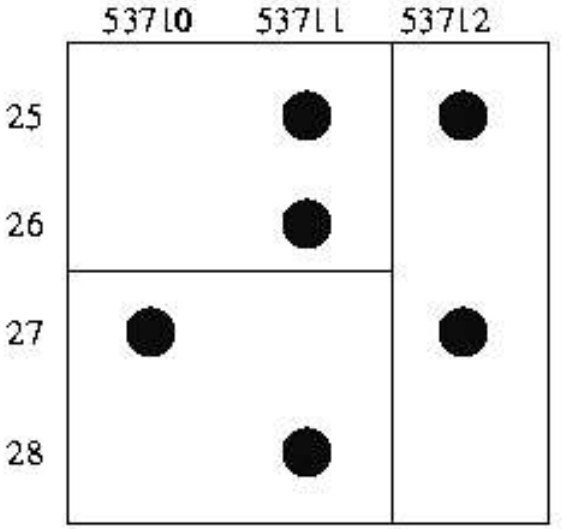
Spatial Representation



(a) Patients



(b) Single – Dimensional



(c) Strict Multidimensional

General-Purpose Quality Metrics

- Discernability Metric

$$C_{DM} = \sum_{EquivClasses\ E} |E|^2$$

- Normalized Average Equivalence Class size

$$C_{AVG} = \left(\frac{TotalRecords}{TotalEquivClasses} \right) / (k)$$

Proposition1

- Every single-dimensional partitioning for quasi-identifiers can be expressed as a strict multidimensional partitioning.
- However, when d>1, there exists a multidimensional partitioning that cannot be expressed as single-dimensional partitioning.
- The problem of finding optimal strict multidimensional partitioning is NP-Hard.

Bounds on Partition Size

- *Allowable Multidimensional cut* :-

If the cut perpendicular along dimension Xi divides partition P into two partitions P1 and P2 such that they have at least K tuples is allowable

- *Allowable Single-dimensional cut* :-

If the cut divides all the regions that it intersects with, in such a way that each resulting region has atleast k tuples then it is allowable

- *Minimal Strict Multidimensional Partitioning* :-

A set S of allowable cuts is minimal partitioning for P if there does not exist a multidimensional allowable cut for P given S

- *Minimal Single-dimensional Partitioning* :-

A set S of allowable cuts is minimal partitioning for P if there does not exist a single-dimensional allowable cut for P given S

Bounds on Partition Size contd.

Theorem 1

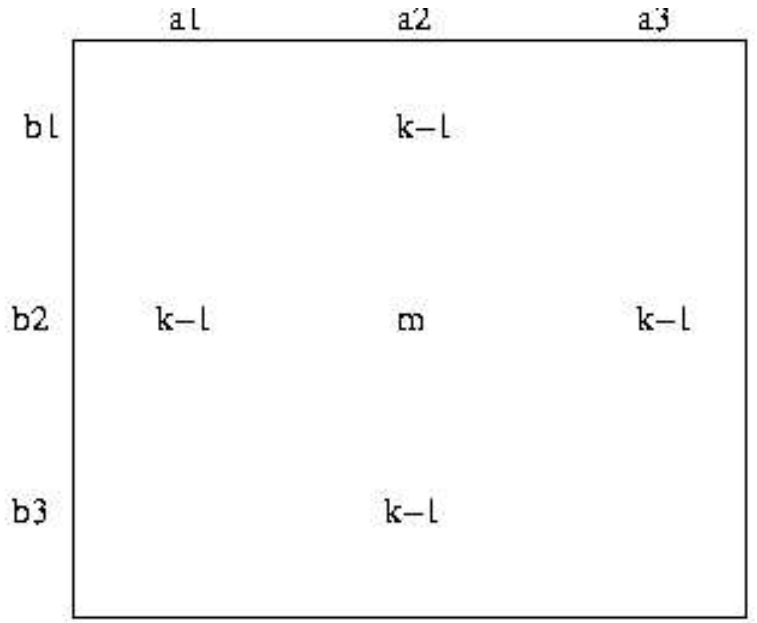
- The maximum number of points contained in any region Ri is $2d(k - 1) + m$

Where,

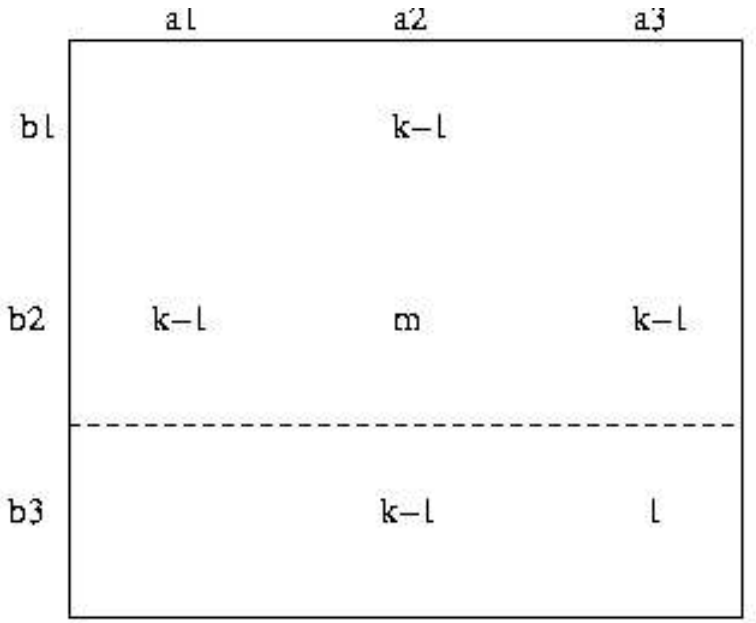
- R1,...., Rn denote the set of regions induced by a minimal strict multidimensional partitioning for multiset of points P

- 'm' is the maximum number of copies of any distinct point in P

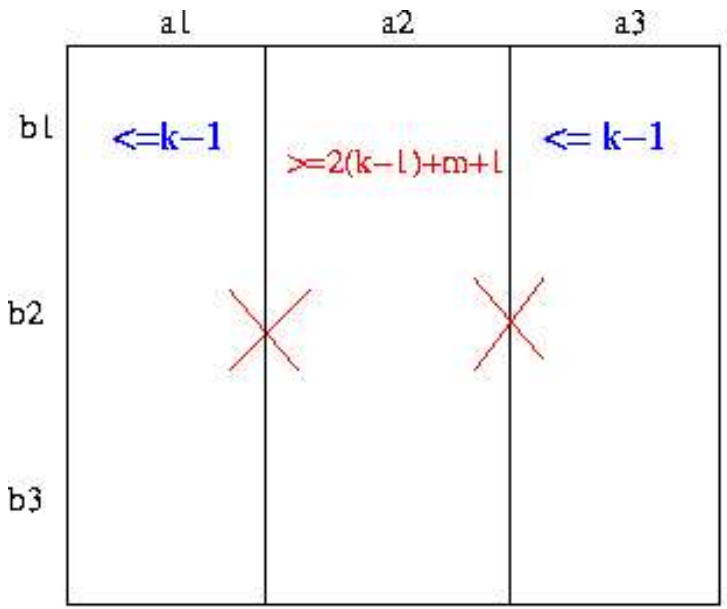
Bounds on Partition Size contd.

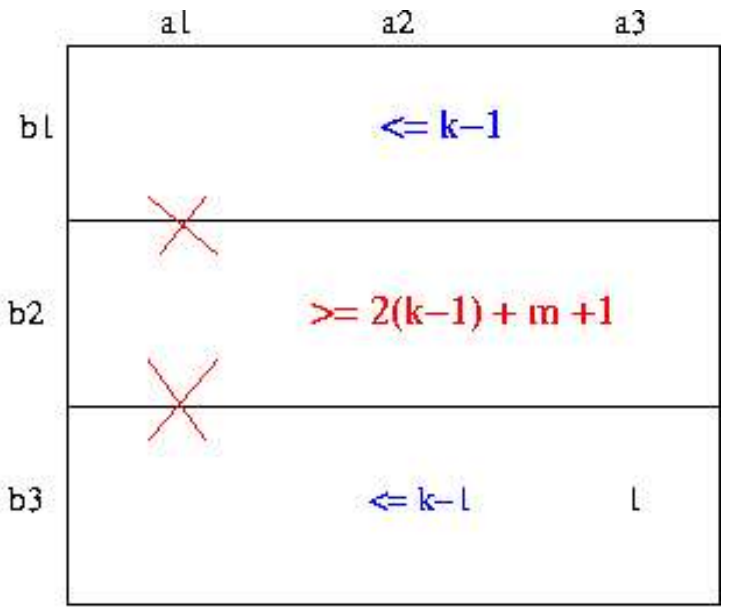


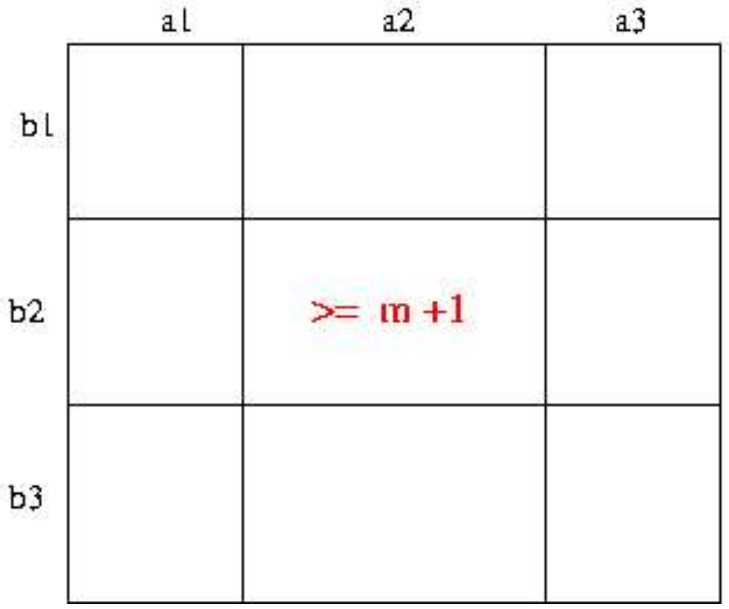
(a) A set of points for which there is no allowable set



(b) Adding a single point produces an allowable cut


a1
a2
a3
b1
b2
b3
<=k−1
>=2(k−1)+m+1
<= k−1


a1
a2
a3
b1
b2
b3
<= k−1
>= 2(k−1) + m +1
<= k−1
1


a1
a2
a3
b1
b2
b3
>= m +1

The Greedy Partitioning Algorithm

Anonymize(partition)
if (no allowable multidimensional cut for partition)
return ø : partition → summary
else
dim ← choose dimension()
fs ← frequency_set(partition, dim)
splitVal ← find median(fs)
lhs ← {t ∈ partition : t:dim ≤ split}
rhs ← {t ∈ partition : t:dim > split}
return Anonymize(rhs) ∪ Anonymize(lhs)

Scalability

- The main issue is finding median of an attribute within a partition when size of table is very large
- Frequency set of the attribute for that partition can be used to calculate the median.
- These sets are much smaller than original table and we can assume that they fit into memory
- In the worst case we need to sequentially scan the the database twice and write it once.

Workload-Driven Quality

- Workload may consist of building a data mining model or answering a set of aggregate queries.
- Ability to answer aggregate depends on the summary statistics provided and the extent to which predicates match range boundaries of data.
- They consider releasing two summary statistics:
  - Range Statistic(R): allows calculation of MIN and MAX aggregates
  - Mean Statistic(M): allows computation of AVG and SUM aggregates

An Example Showing Multiple Summary Statistics

Age(R)	Age(M)	Sex(R)	Zipcode(R)	Disease
[25 – 26]	25.5	Male	53711	Flu
[25 – 27]	26	Female	53712	Hepatitis
[25 – 26]	25.5	Male	53711	Brochitis
[27 – 28]	27.5	Male	[53710 – 53711]	Broken Arm
[25 – 27]	26	Female	53712	AIDS
[27 – 28]	27.5	Male	[53710 – 53711]	Hang Nail

Query 1

```
SELECT AVG(Age)
FROM Patients
WHERE Sex = 'Male'
```

Query 2

```
SELECT COUNT(*)
FROM Patients
WHERE Sex = 'Male'  AND Age ≤ 26
```

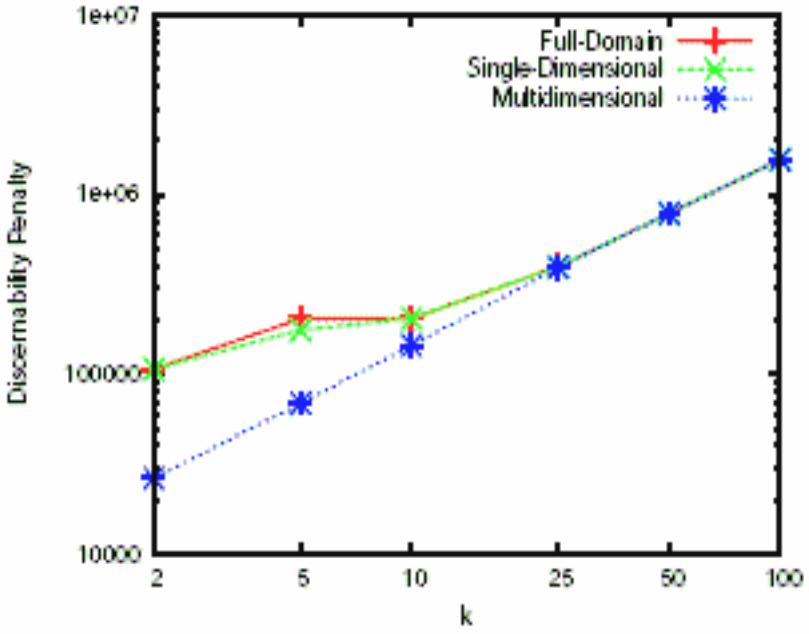
Workload-Driven Anonymization

- In this workload is primarily used for evaluation
- The knowledge of anticipated workload can be integrated into the anonymization algorithm.
- Each query is assigned a weight.
- The algorithm should produce anonymization that reduces the weighted sum of errors caused due to predicates not matching the boundaries of equivalence class.

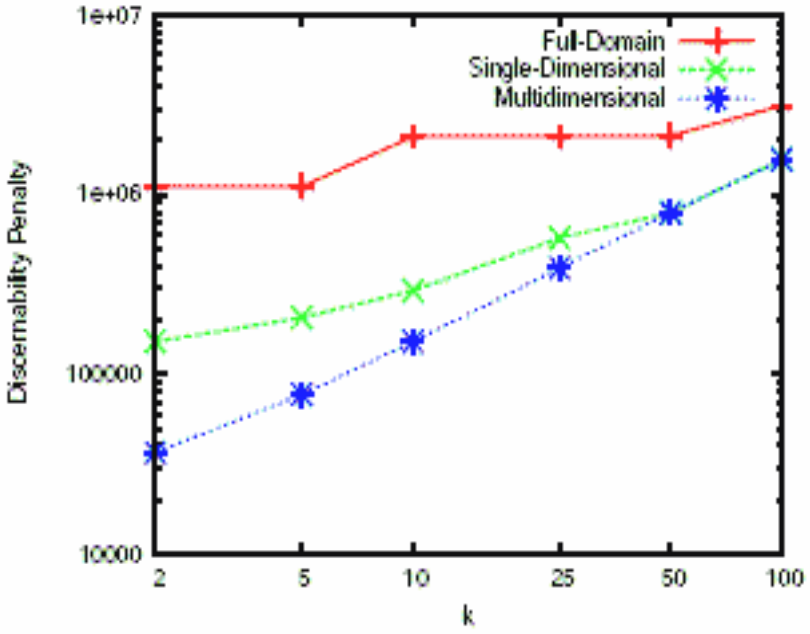
Experimental Evaluation

- They use synthetic data generators that produce two types of distributions for some of their experiments.
- They also used the Adults database from UC Irvine Machine Learning Repository.
- The parameters used for data generation are number of tuples and quasi identifier attributes, cardinality and mean and standard deviation if it is a normal distribution.
- Total no of tuples after configuration was 30162

Experiment 1



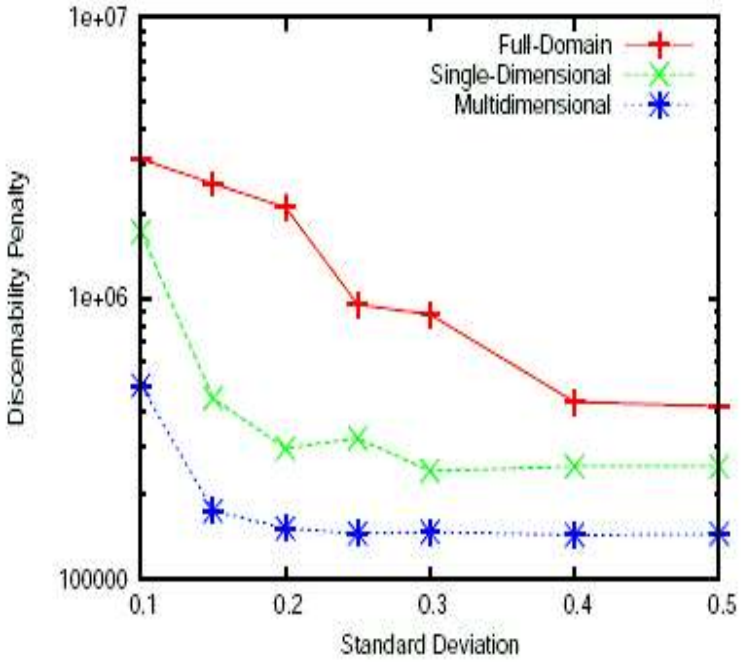
(a) Uniform distribution (5 attributes)



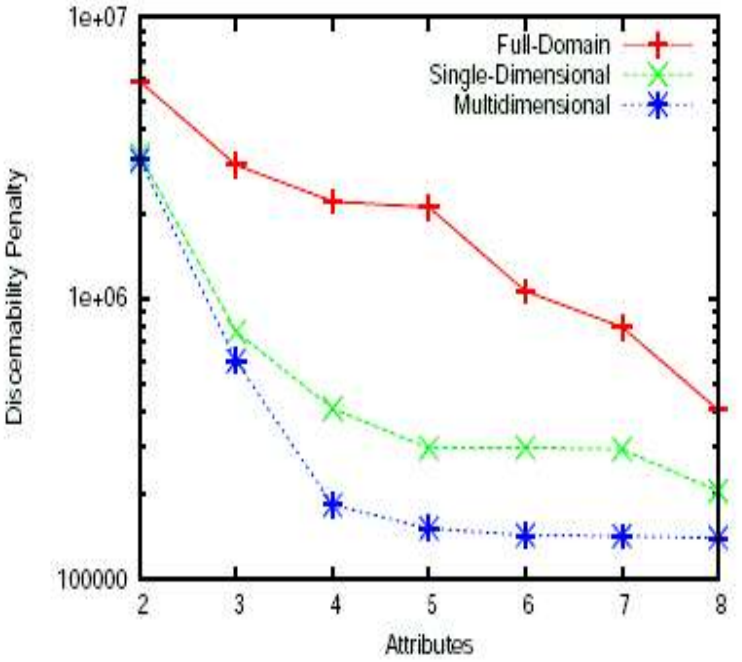
(b) Normal distribution (5 attributes, $\sigma = .2$)

Number of Tuples = 10000 and Attribute Cardinality = 8
For Normal Distribution mean = 3.5

Experiment 2

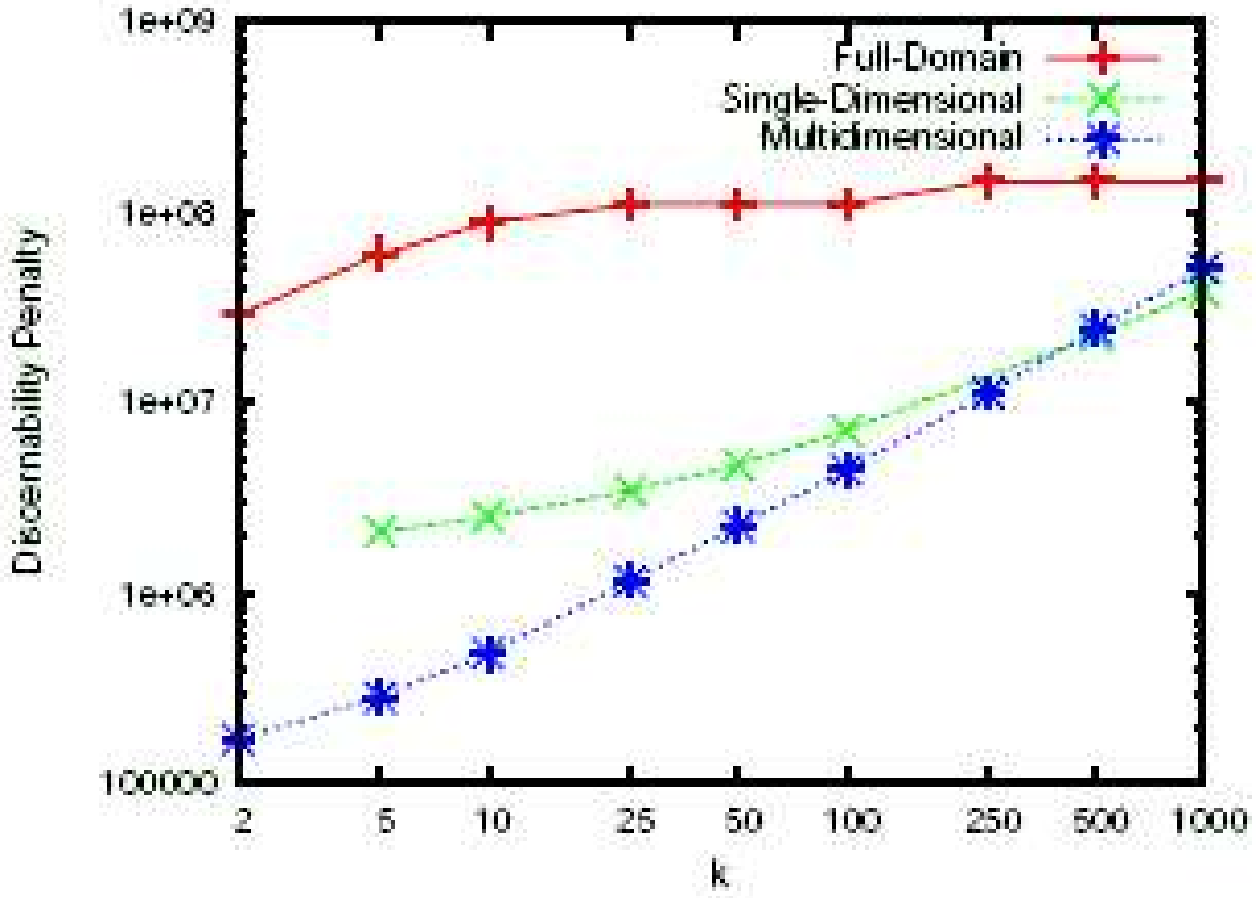


(c) Normal distribution (5 attributes, $k = 10$)

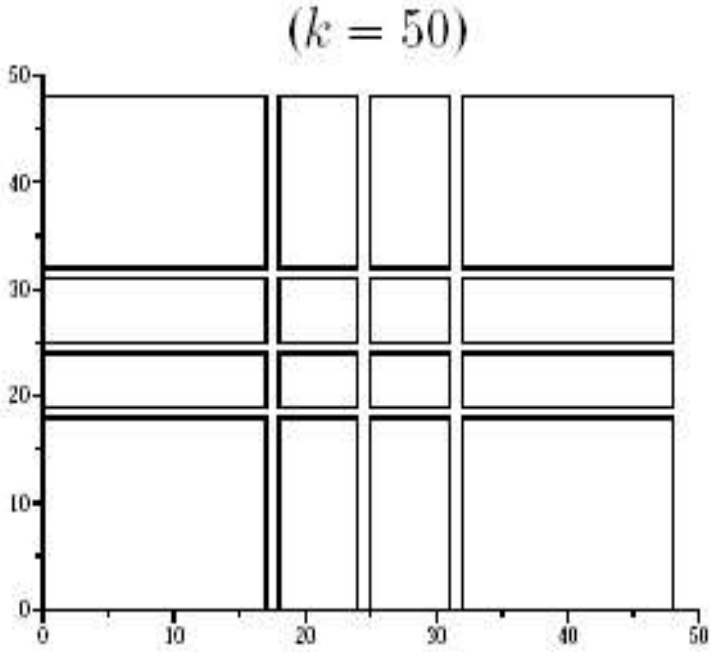


(d) Normal distribution ($k = 10, \sigma = .2$)

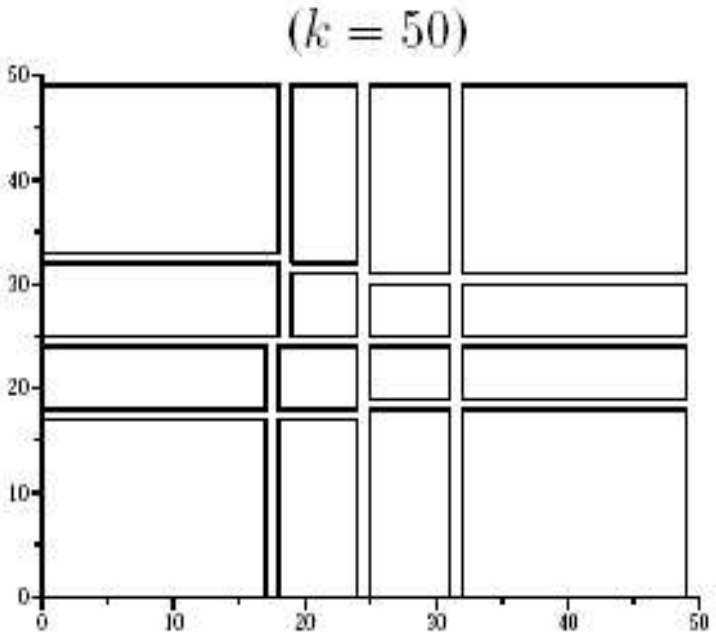
Experiment 3 (Using Adult Database)



Workload Based Quality($\mu = 25$, $\sigma = .2$ cardinality = 50, |T| = 1000)

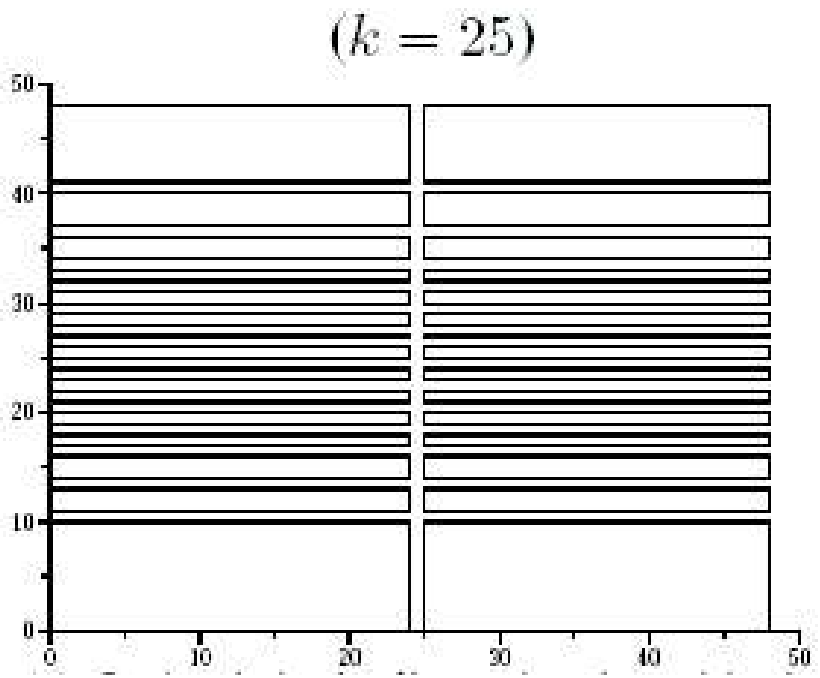


Single-Dimensional

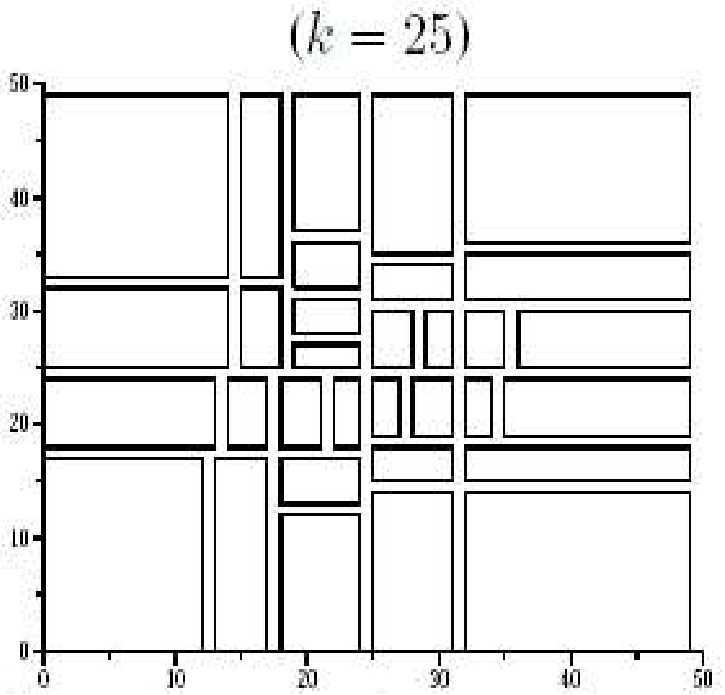


Multidimensional

Workload Based Quality($\mu = 25$, $\sigma = .2$ cardinality = 50, |T| = 1000)

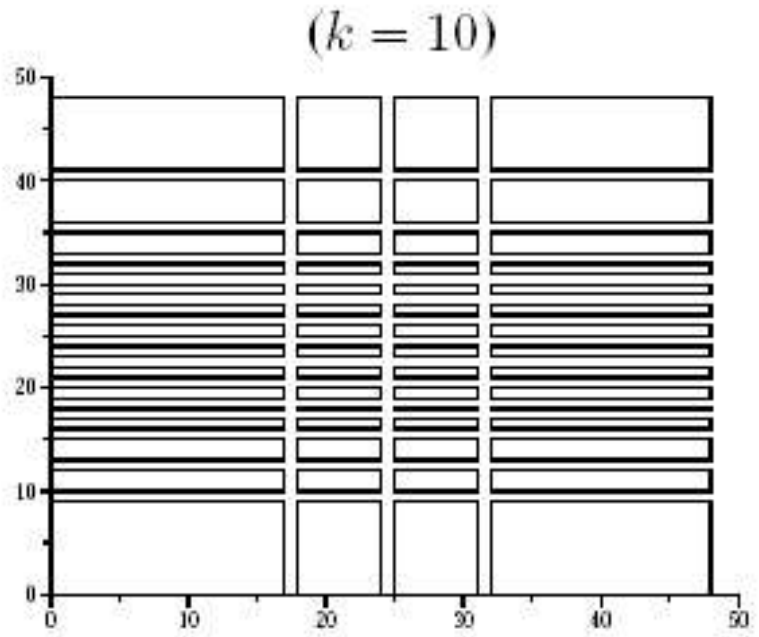


Single-Dimensional

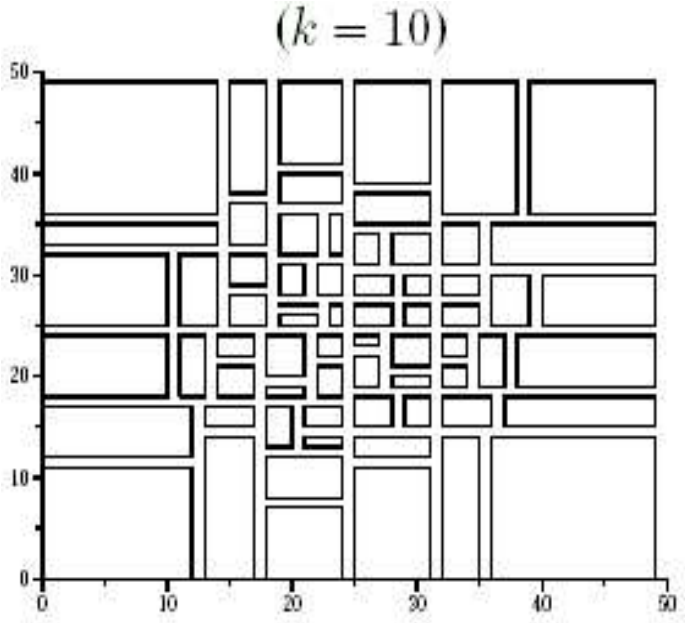


Multidimensional

Workload Based Quality($\mu = 25$, $\sigma = .2$ cardinality = 50, |T| = 1000)



Single-Dimensional



Multidimensional

Errors In Calculation

Queries were of type :-
" SELECT COUNT(*) WHERE {X,Y} = value "

Predicate on X

k	Model	Mean Error	Std. Dev.
10	Single	7.73	5.94
10	Multi	4.66	3.26
25	Single	12.68	7.17
25	Multi	5.69	3.86
50	Single	7.73	5.94
50	Multi	7.94	5.87

Predicate on Y

k	Model	Mean Error	Std. Dev.
10	Single	3.18	2.56
10	Multi	4.03	3.44
25	Single	5.06	4.17
25	Multi	5.67	3.80
50	Single	8.25	6.15
50	Multi	8.06	5.58

xt

Conclusions

- We discussed various models for achieving K-anonymity.
- The greedy algorithm proposed for multidimensional partitioning performs better than other optimal but expensive algorithms.
- This paper gives a better notion of quality based on the workload.
- Multidimensional model performs better for queries involving multiple attributes

References

[1] K. LeFevre, D.DeWitt, and R. Ramakrishnan. Incognito:Efficient full-domain k-anonymity. In ACM SIGMOD 2005.

[2] K. LeFevre, D.DeWitt, and R. Ramakrishnan. Mondrian Multidimensional K - Anonymity. In IEEE ICDE, 2006.

Thank you!
Questions?

Bounds on Quality

- $C_{DMOPT} \geq k * totalrecords$
- $C_{AVGOPT} \geq 1$
- $C_{DM} \leq (2d(k-1)+m) * totalrecords$
- $C_{AVG} \leq (2d(k-1)+m) * total\ records$
- $\frac{C_{DM}}{C_{DMOPT}} \leq \frac{(2d(k-1)+m)}{k}$
- $\frac{C_{AVG}}{C_{AVGOPT}} \leq \frac{(2d(k-1)+m)}{k}$

Bounds on Partition Size contd.

Theorem 3

- The maximum number of points contained in any region R resulting from a minimal single-dimensional partitioning of a multiset of points P is O(|P|)