K-Anonymity & Algorithms

CompSci 590.03

Instructor: Ashwin Machanavajjhala

Duke UNIVERSITY

Announcements

- Project ideas are posted on the site.
  - You are welcome to send me (or talk to me about) your own ideas.

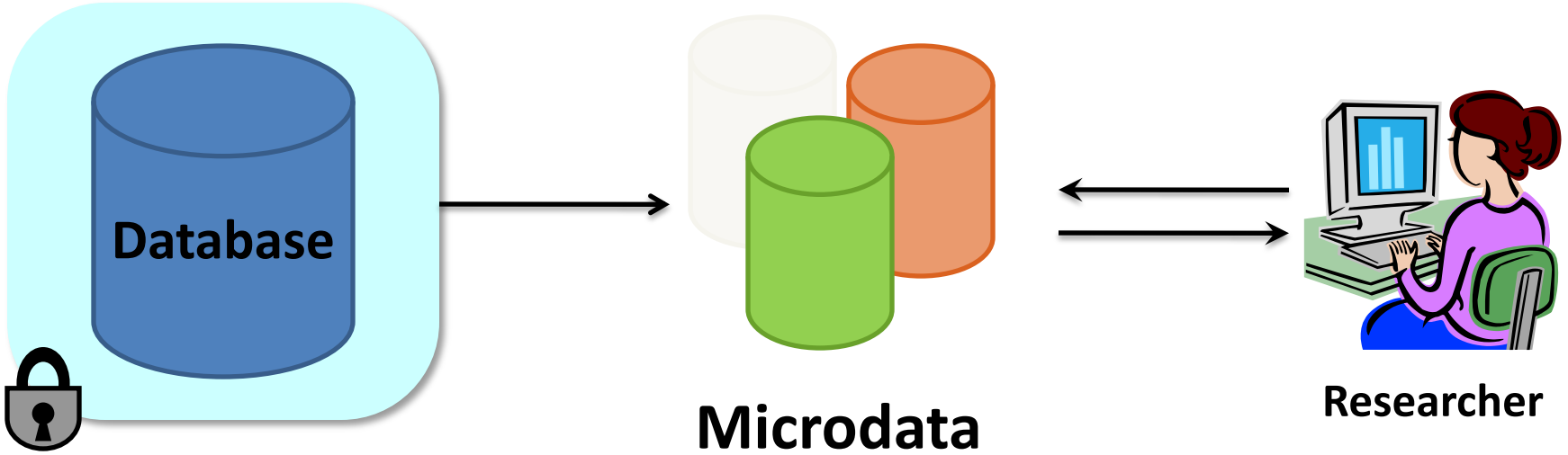
Outline

- K-Anonymity: a metric for anonymity for data publishing
  [Sweeney IJUFKS 2002]

- Algorithms for K-anonymous data publishing
  - Generalization/Suppression
    [Lefevre et al SIGMOD 2006]

  - Curse of Dimensionality
    [Agarwal VLDB 2005]

Offline Data Publishing



Data at the granularity of individuals

Duke UNIVERSITY

Sample Microdata

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0890	14853	50	Indian	Cancer
265-04-1275	14853	55	Russian	Heart
574-22-0242	14850	47	American	Viral
388-32-1539	14850	59	American	Viral
005-24-3424	13053	31	American	Cancer
248-223-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer

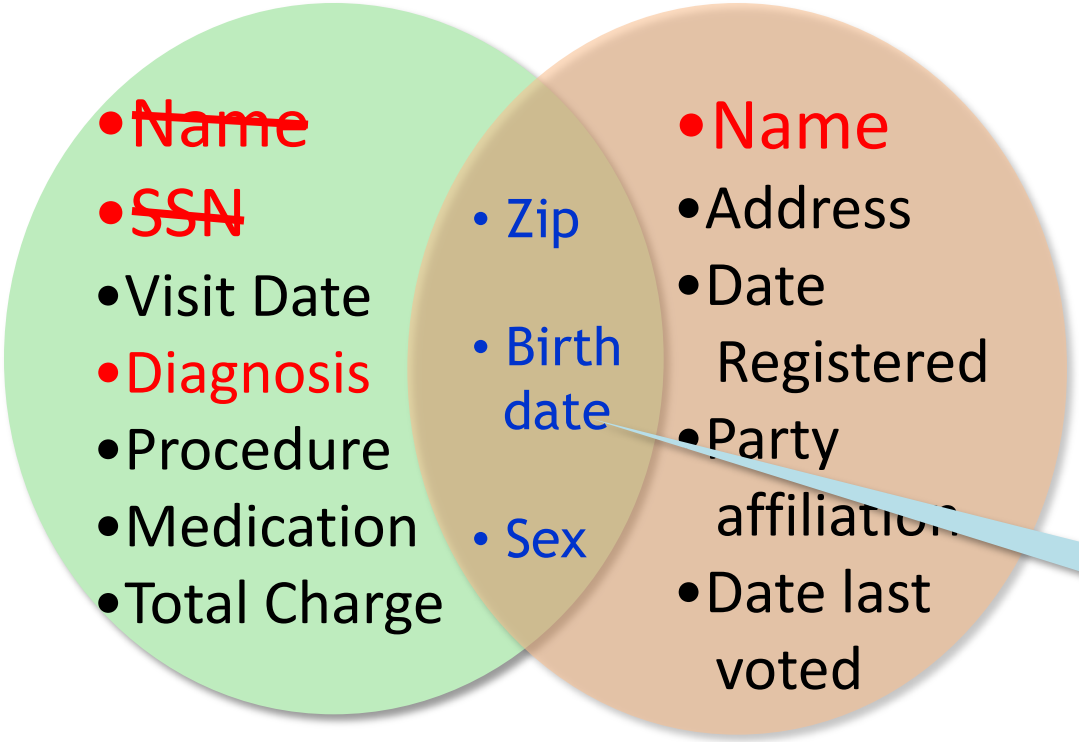
Duke
UNIVERSITY

Removing SSN …

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

Duke
UNIVERSITY

The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]



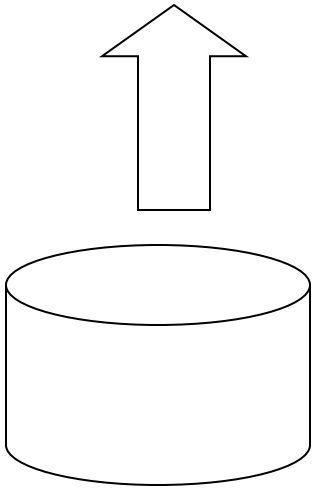
- 87 % of US population **uniquely identified** using ZipCode, Birth Date, and Sex.

Quasi Identifier

Duke UNIVERSITY

Linkage Attacks

Quasi-Identifier



Public Information

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

Duke UNIVERSITY

We saw examples in last class

- Massachusetts governor attack
- AOL privacy breach
- Netflix attack
- Social Network attacks

K-Anonymity

[Samarati et al, PODS 1998]

- Generalize, modify, or distort quasi-identifier values so that no individual is uniquely identifiable from a group of *k*
- In SQL, table T is **k-anonymous** if each

```
SELECT COUNT(*)
FROM T
GROUP BY Quasi-Identifier
```

is **≥ k**

- Parameter k indicates the "degree" of anonymity

Duke UNIVERSITY

Example 1: Generalization (Coarsening)

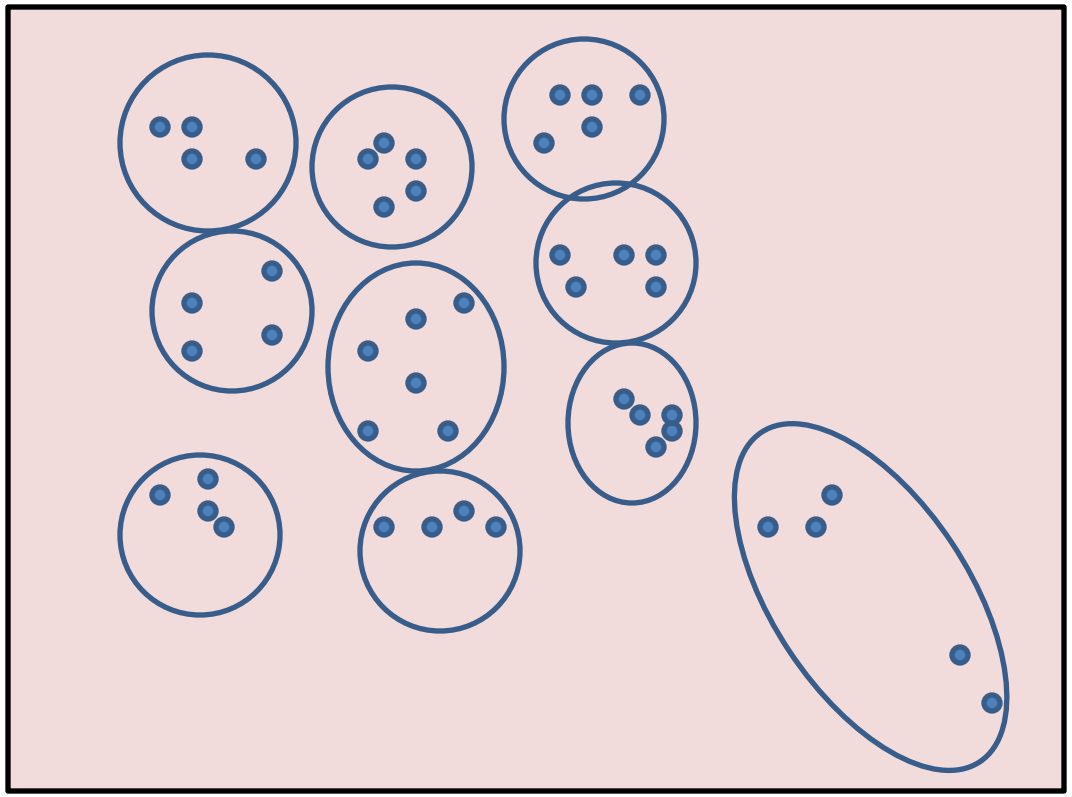
Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053			
13053			
13068			
13068			

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Equivalence Class: Group of k-anonymous records that share the same value for Quasi-identifier attribtutes

Duke UNIVERSITY

Example 2: Clustering



Example 3: Microaggregation

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

Zip	Age	Nationality	Disease
4 tuples Zip code = 130** 23 < Age < 29 Average(age) = 25			2 Heart and 2 Flu
4 tuples Zip = 1485* 47 < Age < 59 Average(age) = 53			1 Cancer, 1 Heart and 2 Flu
4 tuples Zip = 130** 31 < Age < 37 Avergae(age) = 34			All Cancer patients

Duke
UNIVERSITY

K-Anonymity

- Joining the published data to an external dataset using quasi-identifiers results in **at least k records** per quasi-identifier combination.

- What is a quasi-identifier?
  - Combination of attributes (that an adversary may know) that uniquely identify a large fraction of the population.
  - There can be many sets of quasi-identifiers.
    If Q = {B, Z, S} is a quasi-identifier, then Q + {N} is also a quasi-identifier.
  - Need to guarantee k-anonymity against the largest set of quasi-identifiers

Outline

- K-Anonymity: a metric for anonymity for data publishing
  [Sweeney IJUFKS 2002]

- Algorithms for K-anonymous data publishing
  - Generalization/Suppression
    [Lefevre et al SIGMOD 2006]

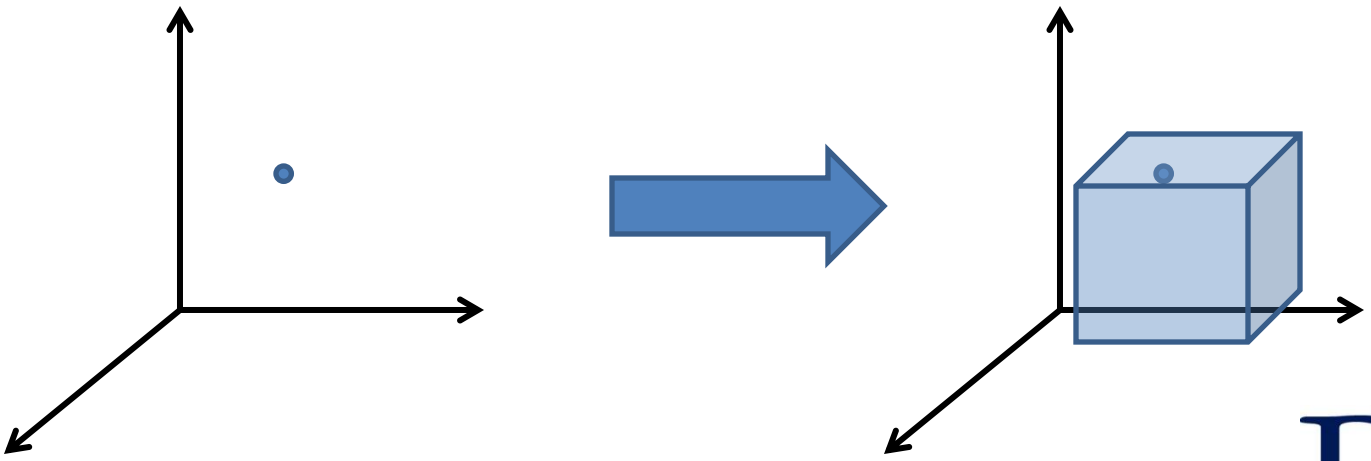
  - Curse of Dimensionality
    [Agarwal VLDB 2005]

Generalization

- Coarsen (or suppress) an attribute to a *more general* value.

- Numeric Values
  - Suppress low significant bits: 12345 -> 1234* -> 123**
  - Ranges: 23 -> [20-25]; (30.5N 20.3E) -> box(30N-31N,20E-22E)

Generation Step



Generalization

- Coarsen (or suppress) an attribute to a *more general* value.

- Categorical Values
  - Domain Generalization Hierarchies
    State-gov occupation → Government occupation → Workclass

Generation Step

Equivalent to suppressing the value

Duke UNIVERSITY

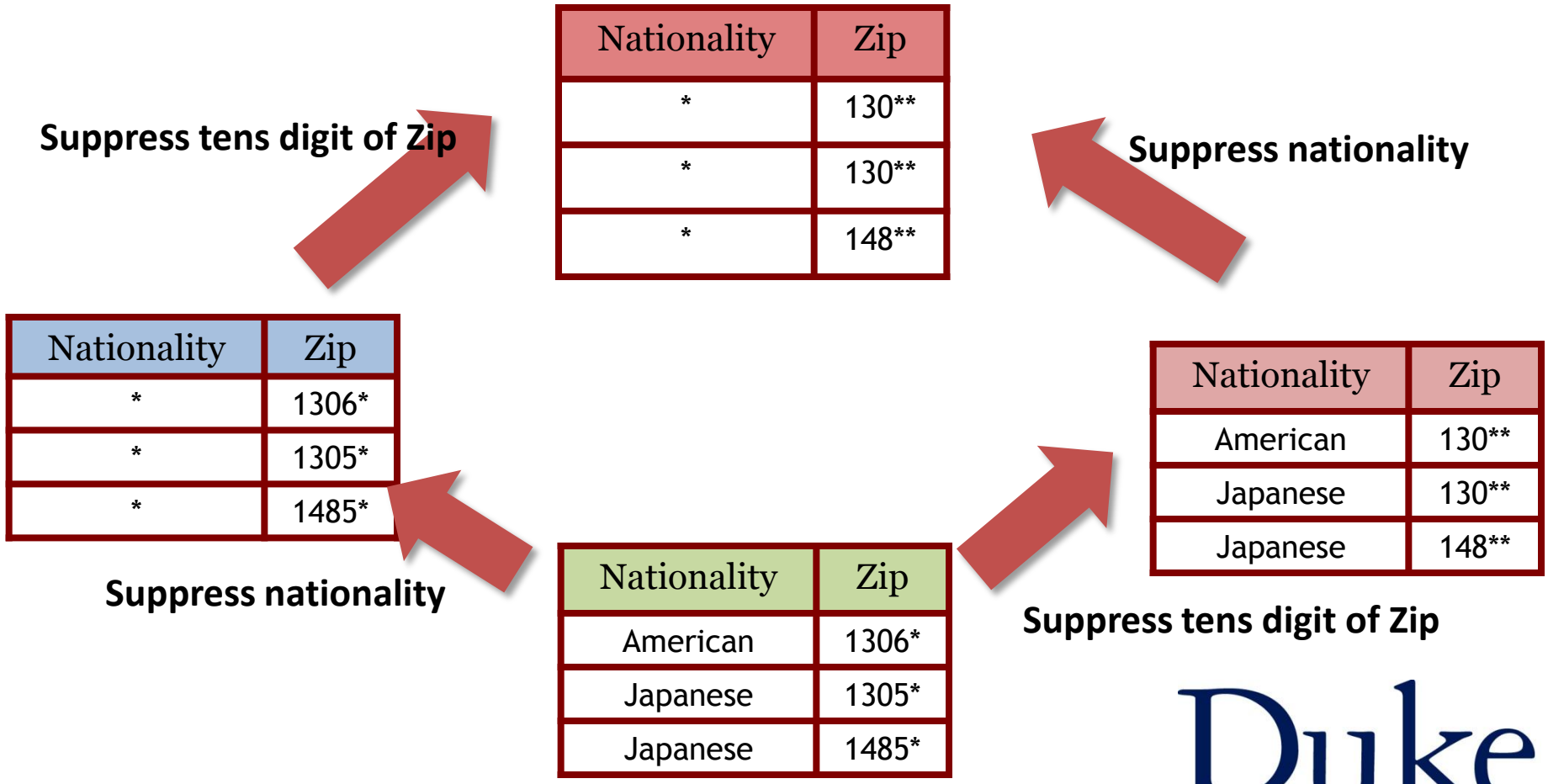
Full Domain vs Local Generalization

- Full Domain:
  Generalize all values in an attribute to the same "level"
  - Every occurrence of 12345 is replaced with 1234* in the database.
  - Answering queries on such datasets is easier.

- Local Generalization:
  Values can be generalized to different levels.
  - 12345 in one tuple may be generalized to 1234*, and in another tuple entirely suppressed.
  - Allows k-anonymous datasets with lesser information loss.

Generalization Lattice

- Generalization step D -> D':
  D' is constructed from D using one generalization step.



Duke UNIVERSITY

Utility: Quantifying error

- Each generalization step introduces error.
- Larger equivalence classes also may lead to more error.

Utility Metrics:

- Average size of equivalence classes
- Number of steps in generalization lattice
- Discernibility metric
  - Assign a penalty to each tuple
  - Penalty depends on how many other tuples are indistinguishable from it

Do not take into account the distribution of values in each equivalence class.

Utility Metrics

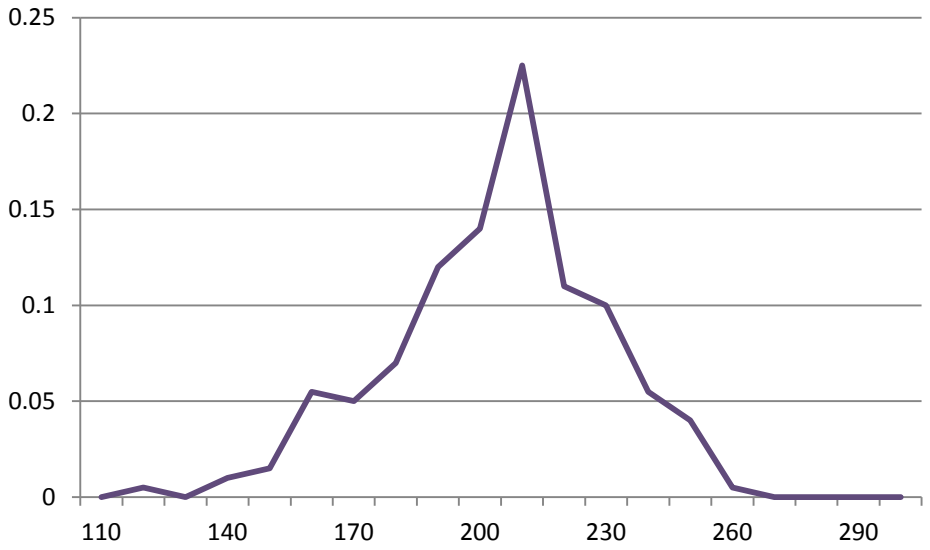
- Classification metric
  - Assign a penalty to each tuple t:
    - If t's sensitive value == majority sensitive value in the group: Penalty = 0
    - Otherwise: Penalty = size of equivalence class

Does not take into account the distribution of the quasi-identifier attributes.

- Information Loss
  - Penalty for each tuple = 1 - 1/ # values that can generalize to that tuple
  - E.g., Penalty (14850, 47) = 1 – 1 /1 = 0
  - Penalty(1485*, [40-50]) = 1 – 1 / (10*10) = .99

Empirical Distribution

- P(X=x) = fraction of tuples in the data with value x.

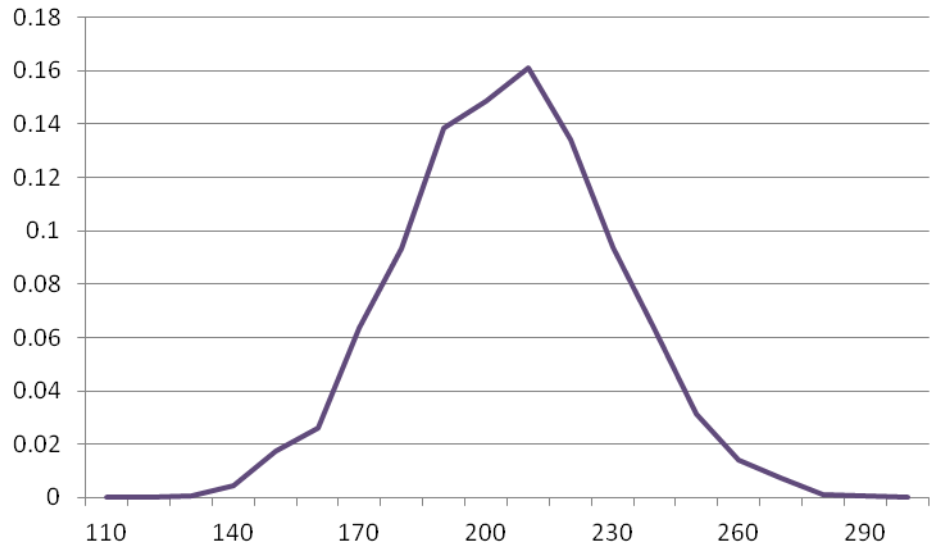


200 weights drawn from a normal distribution with mean 200 and sd 25.

Duke UNIVERSITY

Empirical Distribution

- P(X=x) = fraction of tuples in the data with value x.



2000 weights drawn from a normal distribution with mean 200 and sd 25.

Duke UNIVERSITY

Utility Metrics

KL-Divergence:

- Suppose records were sampled from some multi-dimensional distribution F
  - ***iid*** (identically and independently distributed)
- Given a table, we can estimate F with the empirical distribution F'

F'(14850, 47, American)

= fraction of tuples in the database with
Zip = 14850 AND Age=47 AND Nationality = American

Utility Metrics

KL-Divergence:

- Similarly, given a k-anonymous table, we can compute the empirical distribution $F'_{\text{k-anon}}$

$F'_{\text{k-anon}}(14850, 47, \text{American})$

$= 1/N * (\Sigma_{\text{equivalence class C}} P[(14850, 47, \text{American}) \text{ in C}] * |C|)$

Example

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

F'(13053, 37, Indian) = 1/12

Duke UNIVERSITY

Example

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

$F'_{k\text{-anon}}(13053, 37, \text{Indian}) =$

$= 1/12 (|C_3| * P[(13053, 37, \text{Indian}) \text{ in } C_3])$

$= 1/12 * 4 * 1/(100*10)$

Duke UNIVERSITY

Utility Metrics

Distance between F' and $F'_{k\text{-anon}}$ is a measure of the error due to anonymization

KL-Divergence:

$$\sum_x p(x) \; log \frac{p(x)}{p_{anon}(x)}$$

where p(x) is estimated using the empirical distribution F', and
$p_{anon}(x)$ is estimated using $F'_{k\text{-anon}}$

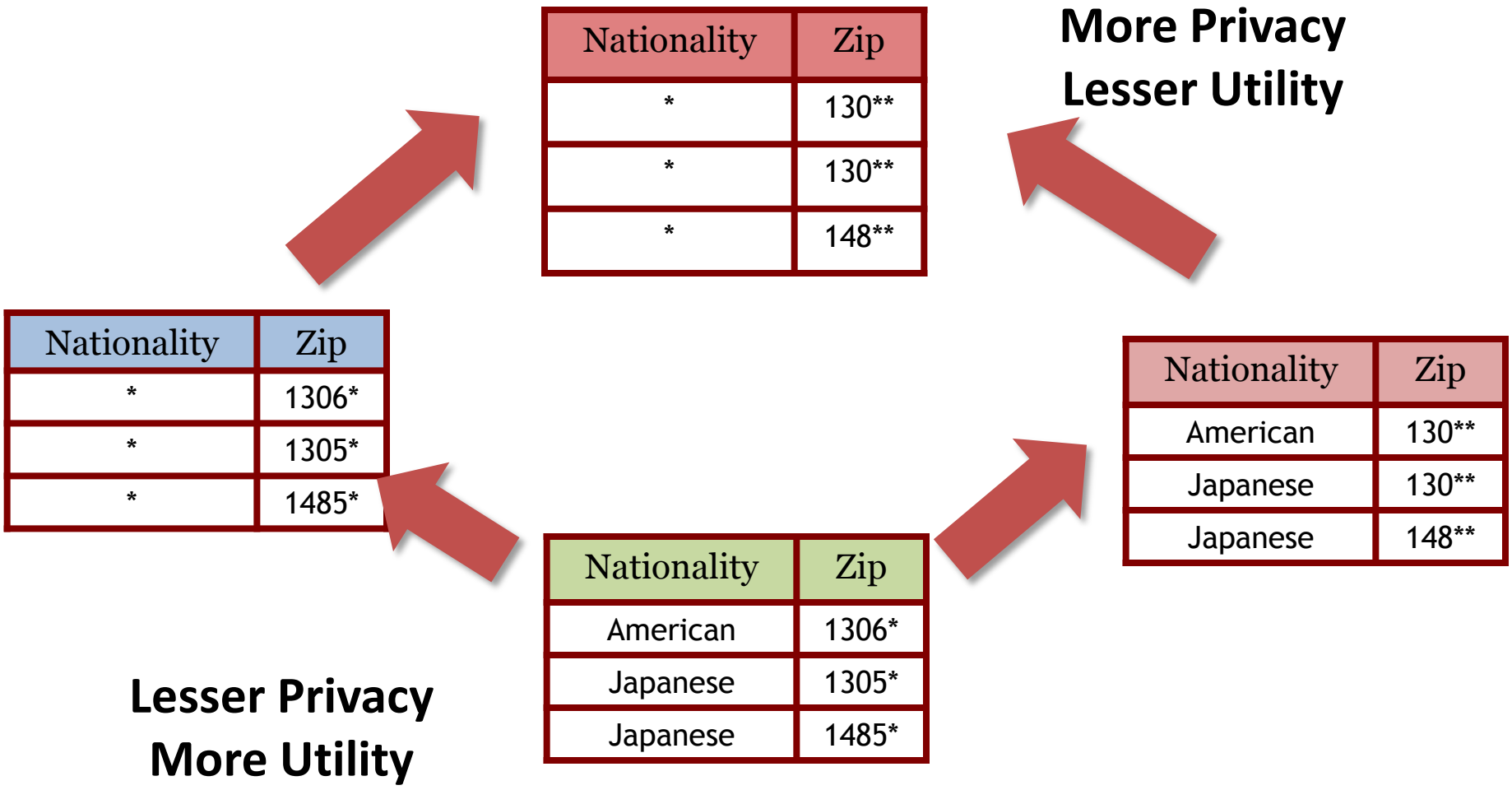
K-Anonymization Problem

Given a table D, find a table D' such that

- D' satisfies the k-anonymity condition
- D' has the maximum utility (minimum information loss)

- NP-Hard [Meyerson & Williams, PODS 2004]
  - Reduction from the k-dimensional matching problem.
  - There is a *log k* approximation algorithm for some utility metrics.

Monotonicity



Duke UNIVERSITY

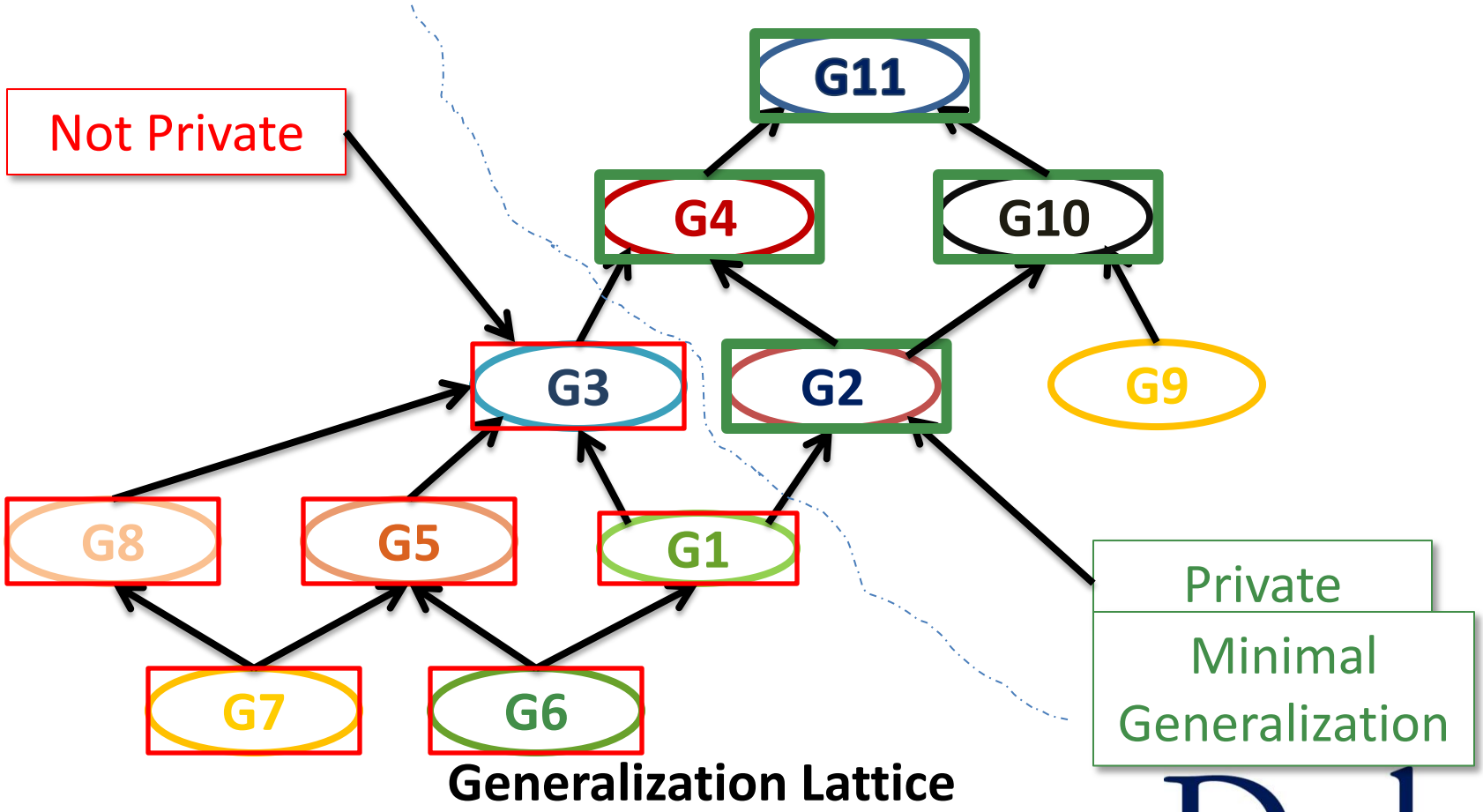
Monotonicity

- In a single generalization step D -> D', new equivalence classes are created by merging existing equivalence classes.

- If D satisfies k-anonymity, then D' also satisfies k-anonymity
  - Equivalence classes are only becoming bigger.

- D' has lesser utility than D
  - Intuitively true: more information is hidden in D'
  - Can be formally shown for all the utility metrics discussed.

Pruning using Monotonicity

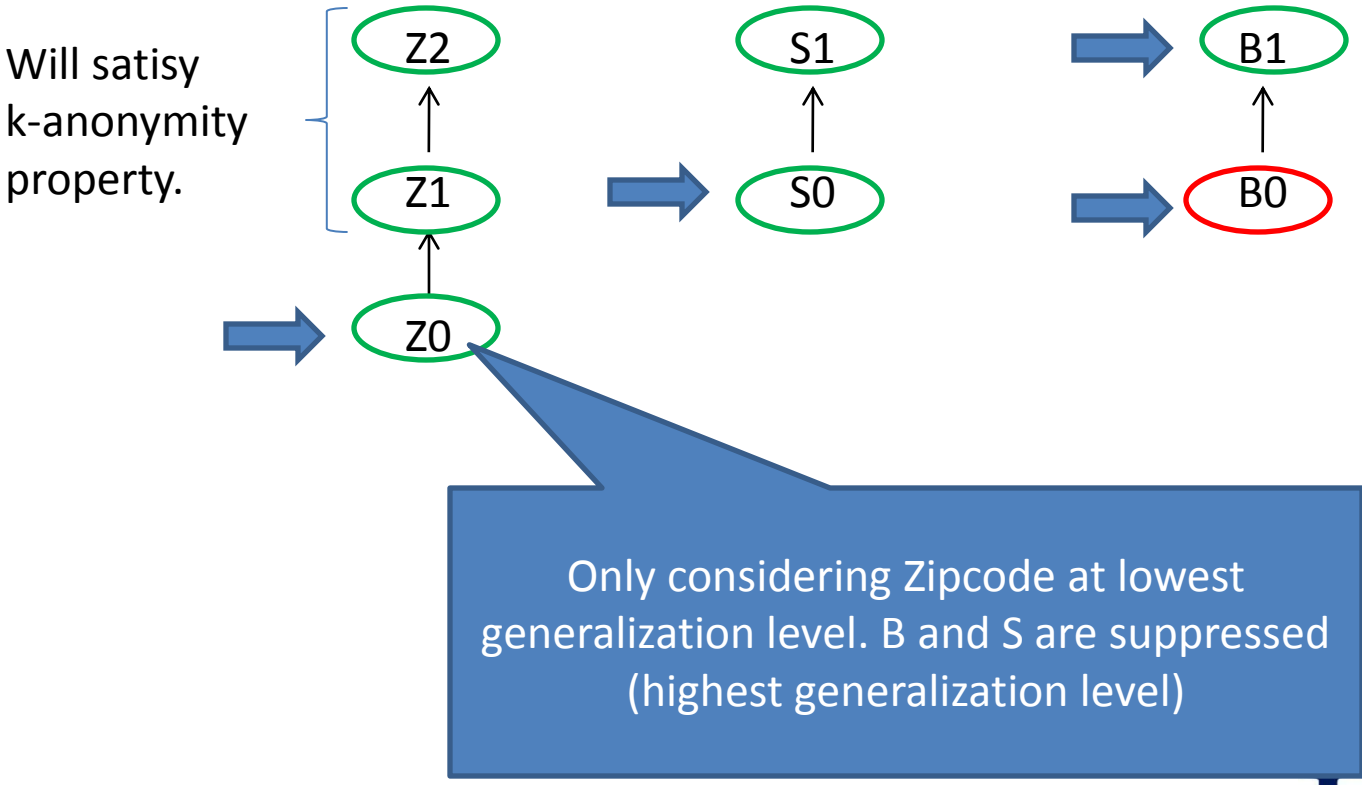


Generalization Lattice

Duke UNIVERSITY

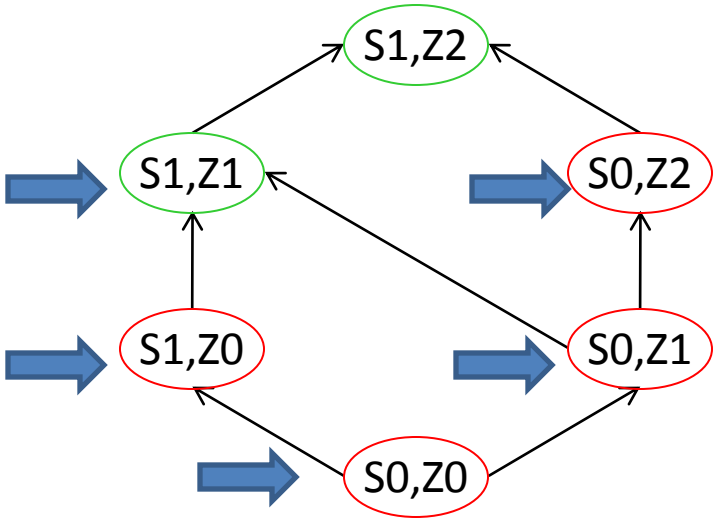
Basic Incognito Algorithm

- Step 1: Start with 1 dimensional quasi-identifier. Start from the bottom of lattice to check when k-anonymity is satisfied.



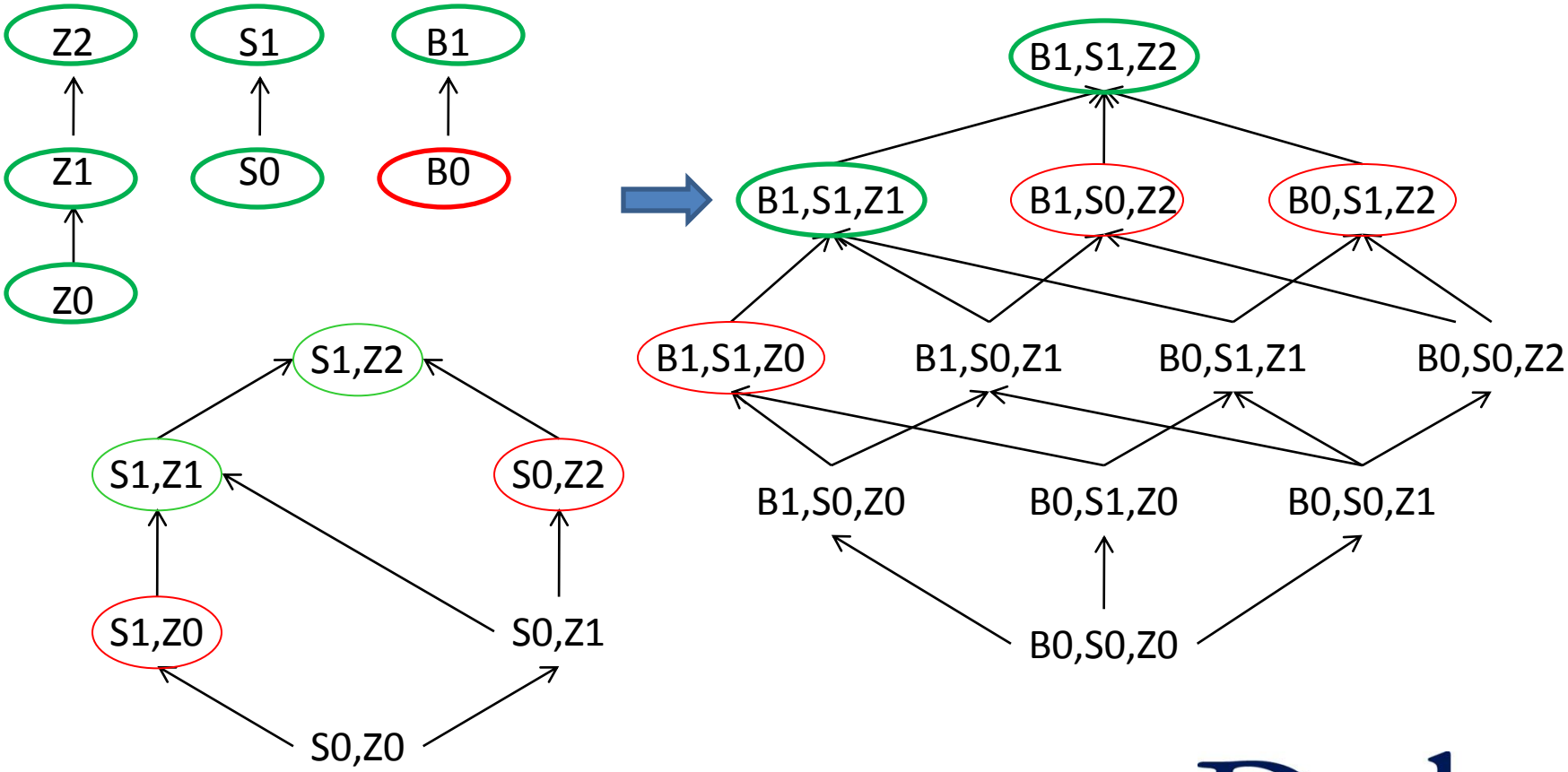
Basic Incognito Algorithm

- Move to 2 dimensional marginals



Basic Incognito Algorithm

- 3-dimensional quasi-identifiers



Duke UNIVERSITY

Summary of Incognito Algorithm

Problem:

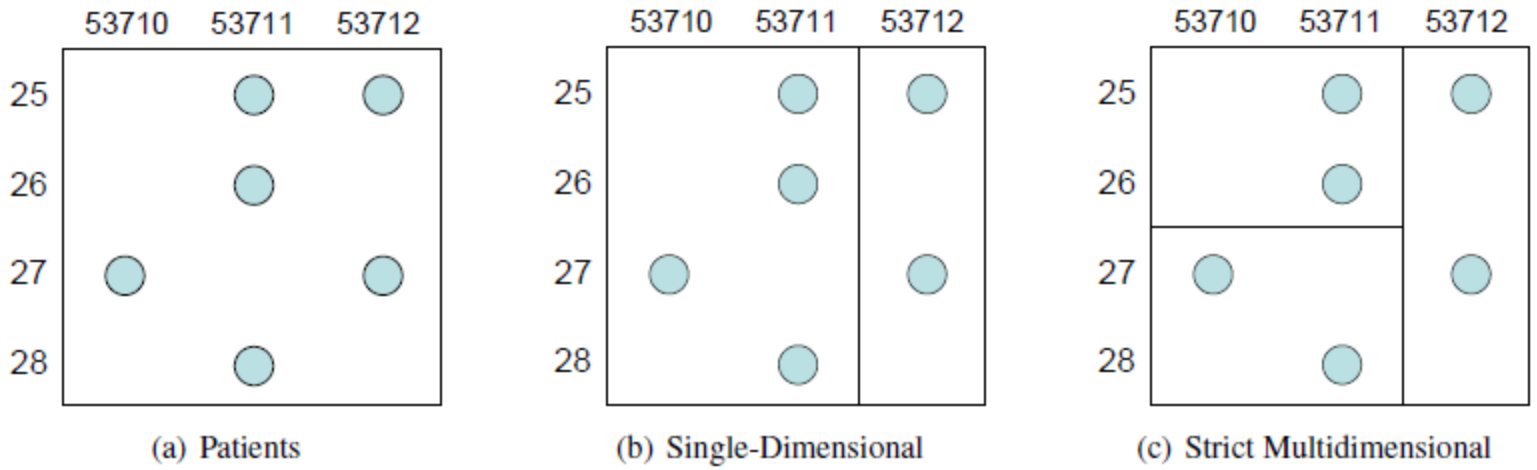
- Amongst all tables that satisfy k-anonymity, find the one that has minimum utility

Solution:

- Generalizations form a Lattice.
- Privacy and Utility are monotonic.
- Only need to find the boundary of "minimal" generalizations that satisfy privacy.
- Lattice can be efficiently pruned using bottom up traversal.
- Checking k-anonymity is efficient (think: precompute counts)

Other K-Anonymity Algorithms

- Mondrian Multidimensional Partitioning [Lefevre et al ICDE 2007]

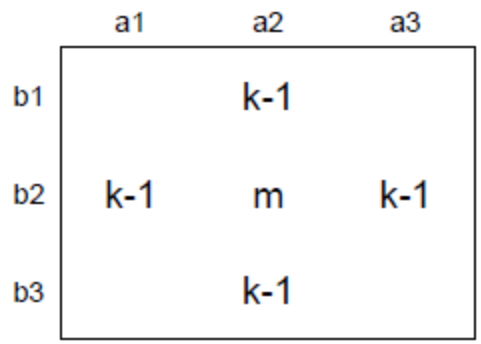


(a) Patients (b) Single-Dimensional (c) Strict Multidimensional

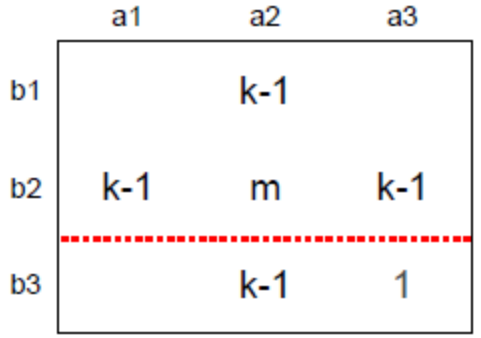
Duke UNIVERSITY

Other K-Anonymity Algorithms

- Mondrian Multidimensional Partitioning



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



(b) Adding a single point produces an allowable cut

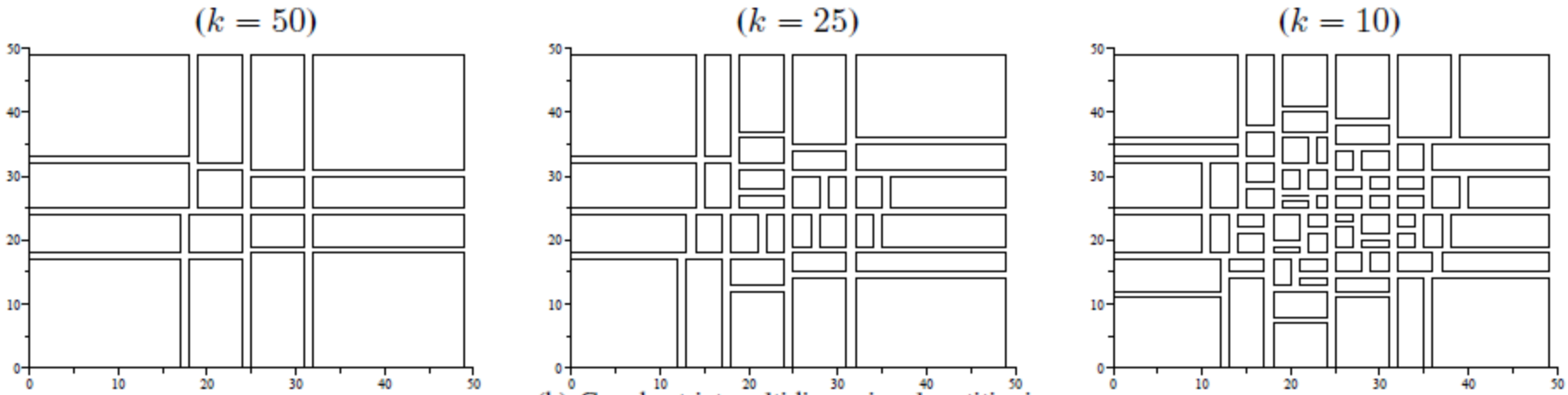
Duke UNIVERSITY

Other K-Anonymity Algorithms

- Mondrian Multidimensional Partitioning
  - Recursive greedy partitioning of the space
  - Partition(region, k)
    1. Choose the best dimension that results in even k-anonymous partition
    2. If possible, partition the region according to that dimension into R1 and R2
    3. Return Partition(R1, k) U Partition(R2, k) // Recurse
    4. If not possible, Return.
  - Workload driven quality metric
    - Utility = error on a set of queries.

Other K-anonymous algorithms

- Mondrian Multidimensional Partitioning

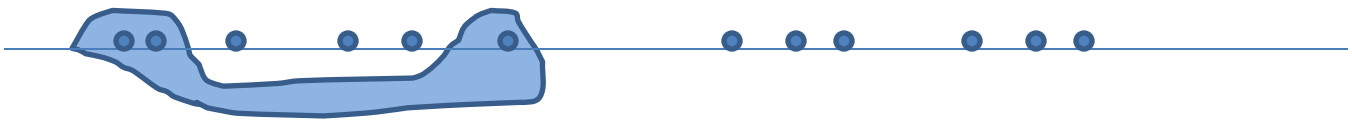


(b) Greedy strict multidimensional partitioning

Duke UNIVERSITY

Other K-anonymous algorithms

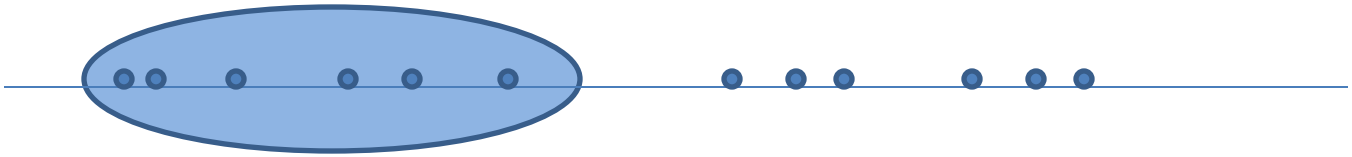
- Hilbert [Ghinita et al VLDB 2007]
  - General k-anonymity is NP-hard
  - Suppose we only have 1 dimensional quasi-identifier?



Never form a group like this.
Contiguous group will have more utility.

Other K-anonymous algorithms

- Hilbert [Ghinita et al VLDB 2007]
  - General k-anonymity is NP-hard
  - Suppose we only have 1 dimensional quasi-identifier?



For k=3, Optimal will never form a group of size >= 6.
Can break it up into 2 groups with better utility.

Other K-anonymous algorithms

- Hilbert [Ghinita et al VLDB 2007]
  - General k-anonymity is NP-hard
  - Suppose we only have 1 dimensional quasi-identifier?

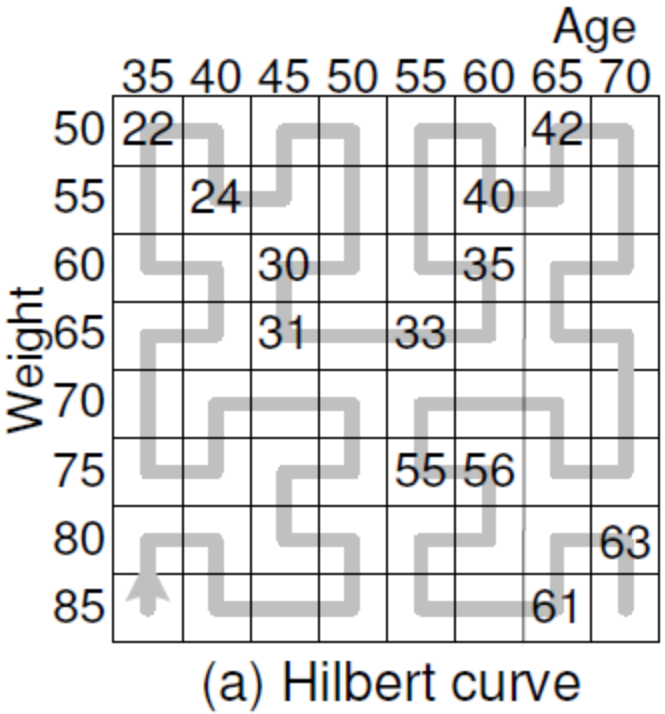
$$Opt(i) = \min_{i-2k<j\leq i-k} (Opt(j) + Opt_I([j+1, i]))$$

Optimal solution for the rest of the points

A group of size at least k and at most 2k-1

Other K-anonymous algorithms

- Hilbert [Ghinita et al VLDB 2007]
  - General k-anonymity is NP-hard
  - But in real datasets, we have multi-dimensional quasi-identifiers.
  - Solution: Map multi-dimensional point to a 1-d point.



(a) Hilbert curve

Duke UNIVERSITY

K-Anonymity by Dissociation

[Terrovitis et al VLDB 2012]

ID	Records
r_1	{itunes, flu, madonna, ikea, ruby}
r_2	{madonna, flu, viagra, ruby, audi a4, sony tv}
r_3	{itunes, madonna, audi a4, ikea, sony tv}
r_4	{itunes, flu, viagra}
r_5	{itunes, flu, madonna, audi a4, sony tv}
r_6	{madonna, digital camera, panic disorder, playboy}
r_7	{iphone sdk, madonna, ikea, ruby}
r_8	{iphone sdk, digital camera, madonna, playboy}
r_9	{iphone sdk, digital camera, panic disorder}
r_{10}	{iphone sdk, digital camera, madonna, ikea, ruby}

(a) Original dataset D

K = 3

		Record chunks		Term chunk
		C_1	C_2	C_T
Cluster P_1 $\lvert P_1\rvert = 5$	r_1	{itunes, flu, madonna}		
	r_2	{madonna, flu}	{audi a4, sony tv}	ikea, viagra,
	r_3	{itunes, madonna}	{audi a4, sony tv}	ruby
	r_4	{itunes, flu}		
	r_5	{itunes, flu, madonna}	{audi a4, sony tv}	

		Record chunk	Term chunk
		C_1	C_T
Cluster P_2 $\lvert P_2\rvert = 5$	r_6	{madonna, digital camera}	
	r_7	{iphone sdk, madonna}	
	r_8	{iphone sdk, digital camera, madonna}	panic disorder,
	r_9	{iphone sdk, digital camera}	playboy, ikea, ruby
	r_{10}	{iphone sdk, digital camera, madonna}	

(b) Anonymized dataset D^A

Duke UNIVERSITY

Curse of Dimensionality

[Beyer et al ICDT 1999]
[Agarwal VLDB 2005]

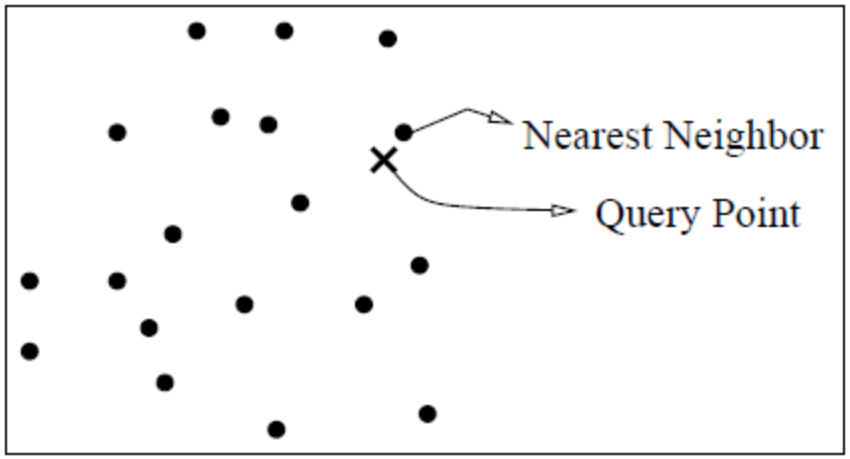


Fig. 1. Query point and its nearest neighbor.

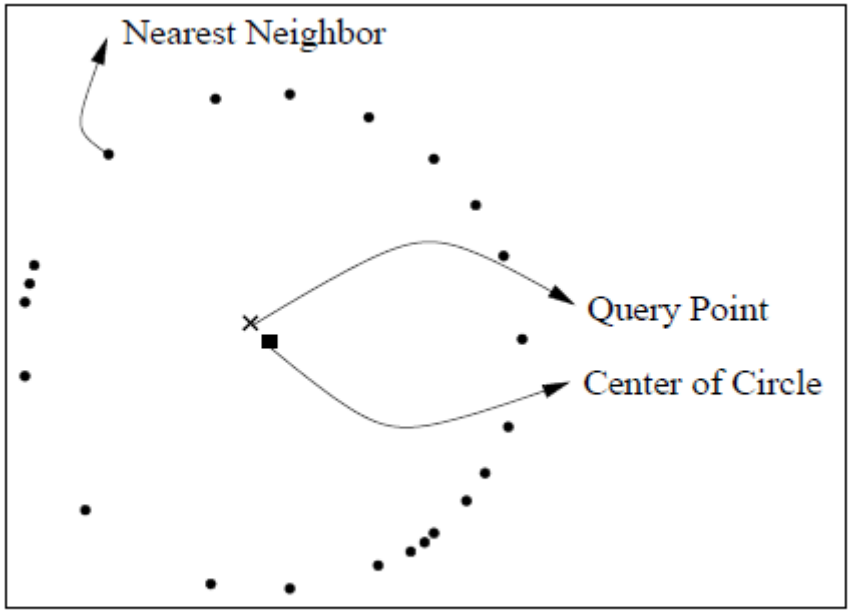


Fig. 2. Another query point and its nearest neighbor.

Next Class

- Ensuring K-Anonymity in Social Networks

References

L. Sweeney, "*K-Anonymity: a model for protecting privacy*", IJUFKS 2002

K. Lefevre, D. Dewitt & R. Ramakrishnan, "*Incognito: Efficient Full Domain K-Anonymization",* SIGMOD 2006

K. Lefevre, D. Dewitt & R. Ramakrishnan, "*Mondrian Multidimensional k-anonymity",* ICDE 2007

G. Ghinita, P. Karras, P. Kalnis & N. Mamoulis, *"Fast Data Anonymization with Low Information Loss"*, VLDB 2007

M. Terrovitis, J. Liagouris, N. Mamoulis & S. Skiadopolous, "*Privacy Preservation by Disassociation*", VLDB 2012

K. Beyer, J. Goldstein, R. Ramakrishnan & U. Shaft, "*When is "nearest neighbor" meaningful?*", ICDT 1999

C. Agarwal, "*On K-Anonymity and the Curse of Dimensionality*", VLDB 2005