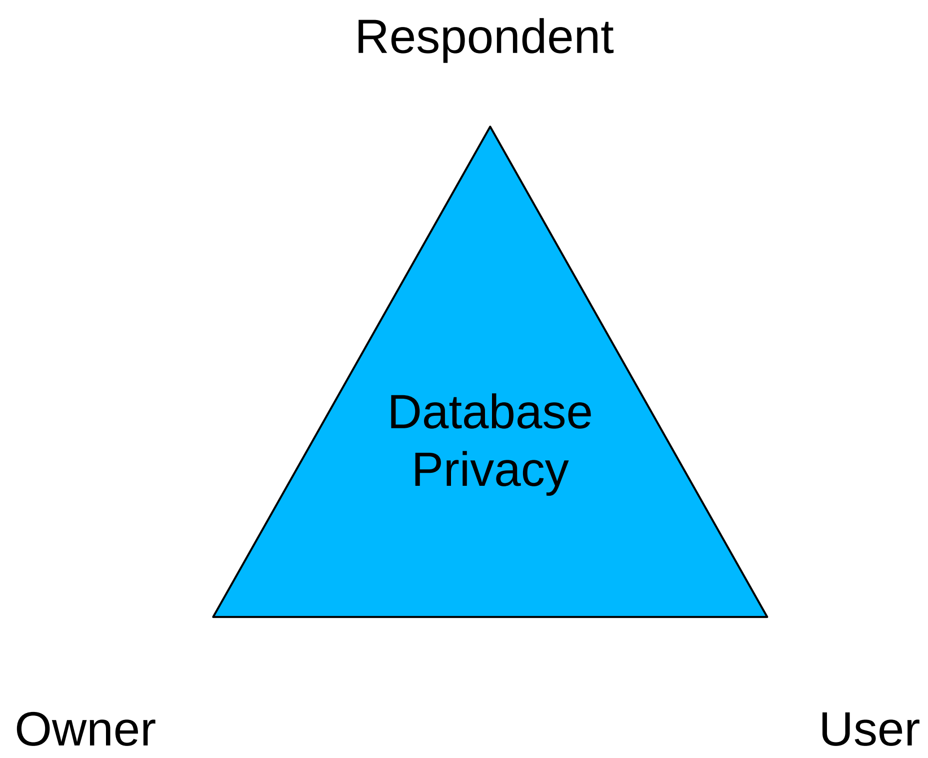
Privacy Preserving Data Publishing

Prof. Ravi Sandhu
Executive Director and Endowed Chair

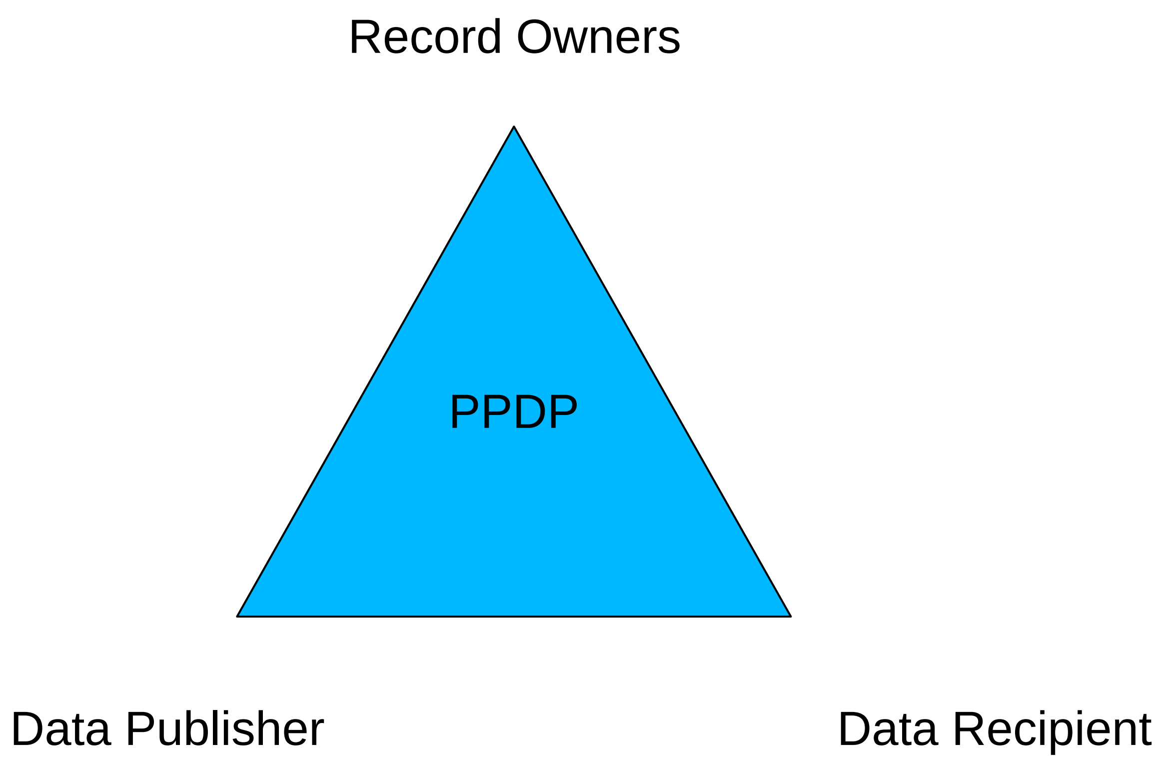
March 29, 2013

ravi.sandhu@utsa.edu
www.profsandhu.com

Domingo-Ferrer 2007

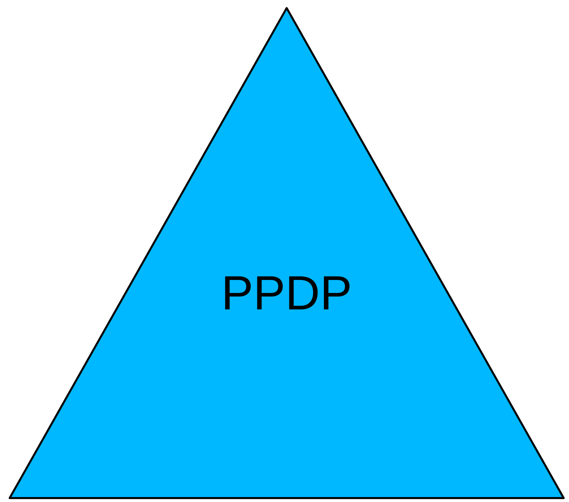


3 independent dimensions


Record Owners
PPDP
Data Publisher
Data Recipient

Fung et al 2010

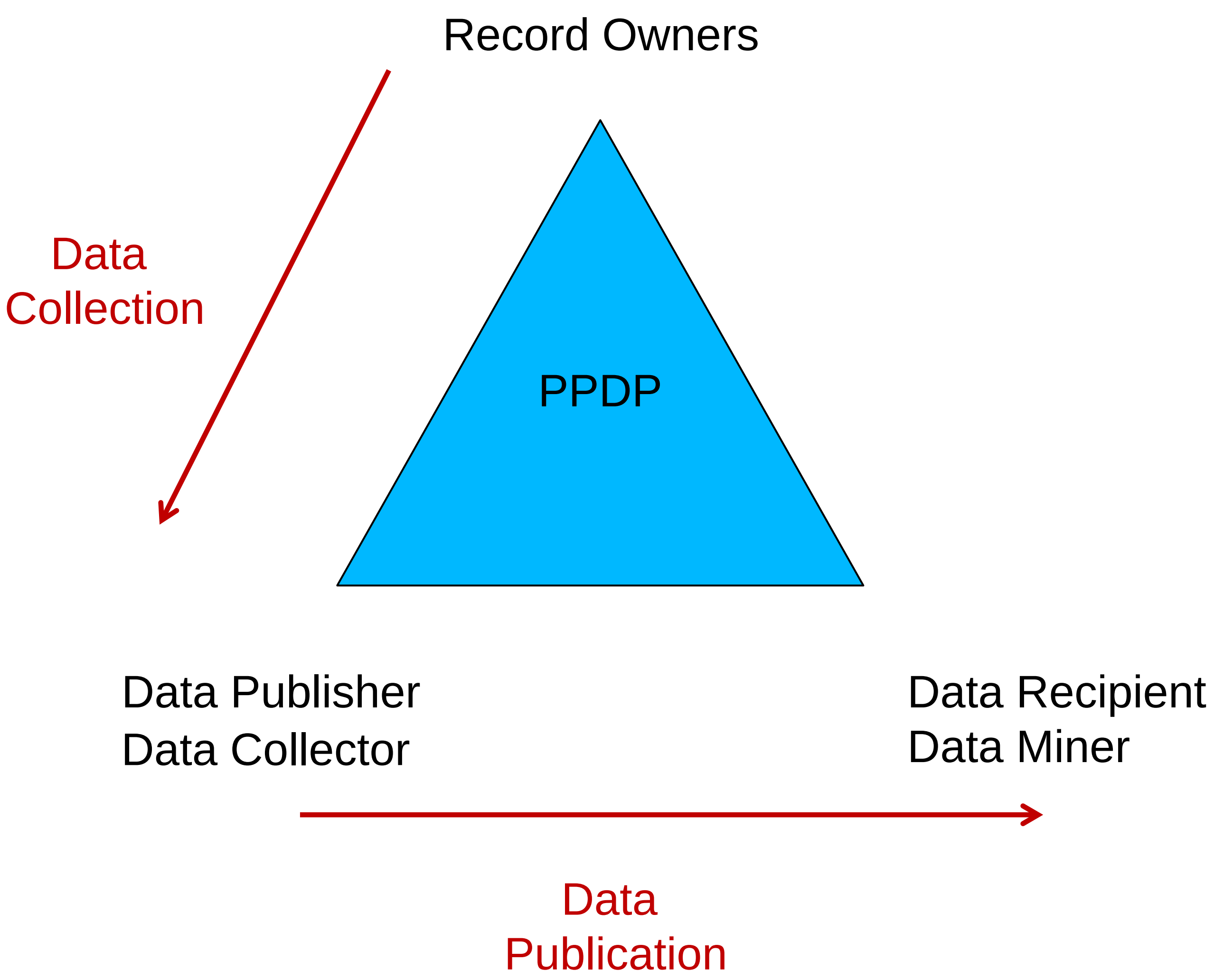
Record Owners



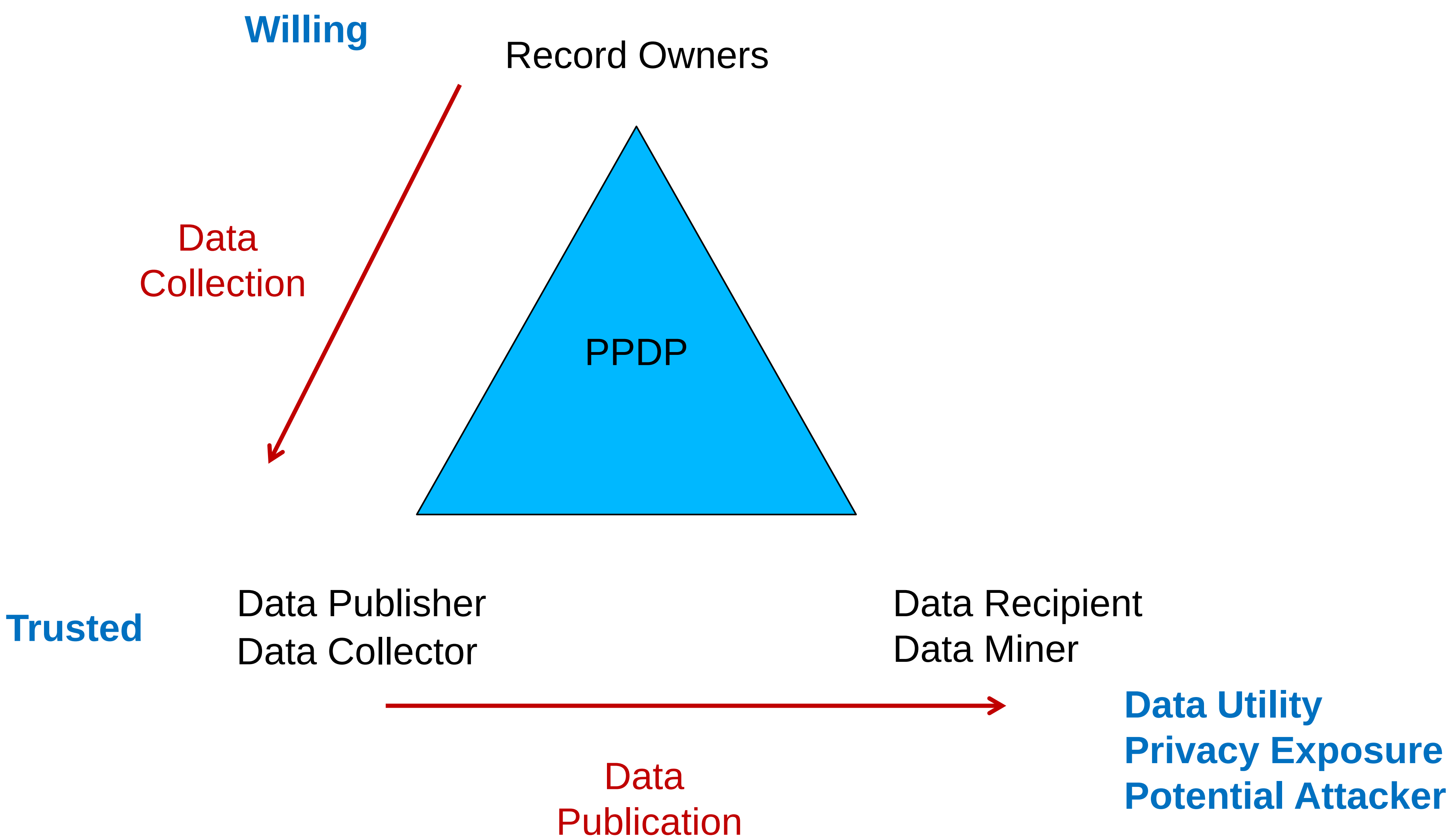
Data Publisher
Data Collector

Data Recipient
Data Miner

Fung et al 2010



Fung et al 2010



Related but not Synonymous

- Privacy preserving data mining (PPDM)
  - How to do data mining when the publisher has modified the data to obscure sensitive information?
  - How to modify the data to obscure sensitive information without loosing ability to data mine?
  - Techniques often tied to data mining task.
  - PPDM is being used even when no data mining as such is being done.

Records in a Table

- Single table
- Each record pertains to a distinct owner (typically)
- 4 kinds of attributes (disjoint):
  - Explicit identifier
  - Quasi identifier (QID)
  - Sensitive attributes
  - Non-sensitive attributes
- Anonymization techniques
  - Modified quasi identifier (QID')
  - Add noise
  - Generate synthetic data "similar" to original

Privacy Definition: Absolute Privacy

- Absolute Privacy, Dalenius 1977
  - Access to published data should not enable the attacker to learn anything extra about any target victim compared to no access to the database, even with the presence of any attacker's background knowledge obtained from other sources.
- Impossible, Dwork 2006
  - Even if published data does not include target victims record attacker can still learn something about target victim from published data and background knowledge.

Privacy Definition: Differential Privacy

- Differential Privacy, Dwork 2006
  - Compare risk to target victim’s privacy with or without presence of target victim’s record in published database.
  - Risk should not substantially increase if the record is included.

Privacy Definition: Uninformative Principle

- Uninformative Principle, Machanavajjhala et al 2006
  - Difference between prior and posterior beliefs is small

Privacy Definition: Linkage Attacks

- Record linkage
- Attribute linkage
- Table linkage

Record Linkage

Table II. Examples Illustrating Various Attacks

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

(c) 3-anonymous patient table

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

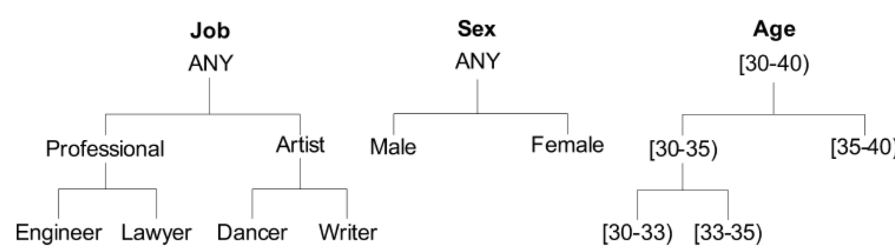


Fig. 3. Taxonomy trees for *Job*, *Sex*, *Age*.

QID

Attribute Linkage

Table II. Examples Illustrating Various Attacks

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

Attribute Linkage

(c) 3-anonymous patient table

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

Table Linkage

(c) 3-anonymous patient table

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

(d) 4-anonymous external table

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Bob	Professional	Male	[35-40)
Cathy	Artist	Female	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Professional	Male	[35-40)
Gladys	Artist	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

Published

Known to be subset of

Public

Probability that Alice is in (c) is 4/5
Probability that Bob is in (c) is 3/4

Privacy Models and Attacks

Table I. Privacy Models

Privacy Model	Attack Model			
	Record Linkage	Attribute Linkage	Table Linkage	Probabilistic Attack
k-Anonymity	✓			
MultiR k-Anonymity	✓			
ℓ-Diversity	✓	✓		
Confidence Bounding		✓		
(α, k)-Anonymity	✓	✓		
(X, Y)-Privacy	✓	✓		
(k, e)-Anonymity		✓		
(ϵ, m)-Anonymity		✓		
Personalized Privacy		✓		
t-Closeness		✓		✓
δ-Presence			✓	
(c, t)-Isolation	✓			✓
ϵ-Differential Privacy			✓	✓
(d, γ)-Privacy			✓	✓
Distributional Privacy			✓	✓

Anonymization Techniques

- Generalization and suppression
- Anatomization and permutation
- Perturbation

Generalization

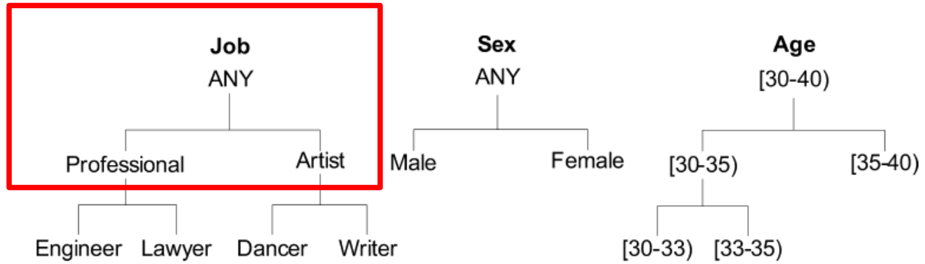


Fig. 3. Taxonomy trees for *Job*, *Sex*, *Age*.

- Full domain generalization
  - Generalize to same level in tree
- Subtree generalization
- Sibling generalization
- Cell generalization
  - Local recoding versus global recoding for above
- Multi-dimensional generalization

Generalization

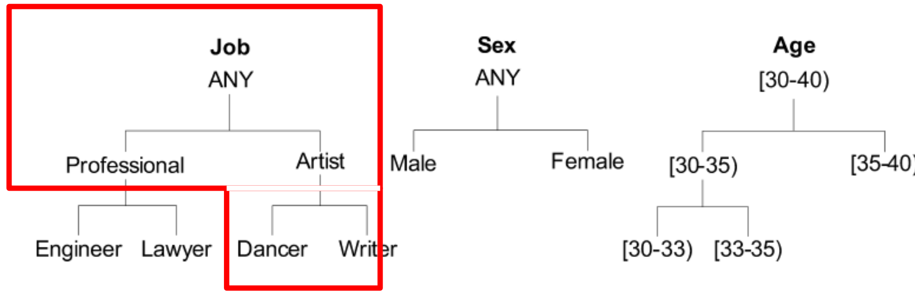


Fig. 3. Taxonomy trees for *Job*, *Sex*, *Age*.

- Full domain generalization
  - Generalize to same level in tree
- Subtree generalization
- Sibling generalization
- Cell generalization
  - Local recoding versus global recoding for above
- Multi-dimensional generalization

Generalization

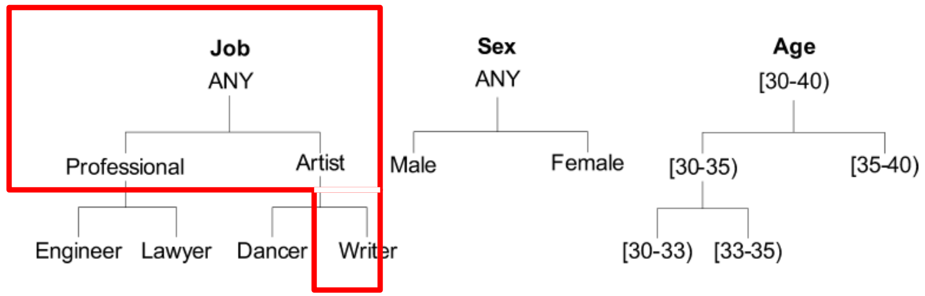


Fig. 3. Taxonomy trees for *Job*, *Sex*, *Age*.

- Full domain generalization
  - Generalize to same level in tree
- Subtree generalization
- Sibling generalization
- Cell generalization
  - Local recoding versus global recoding for above
- Multi-dimensional generalization

Generalization

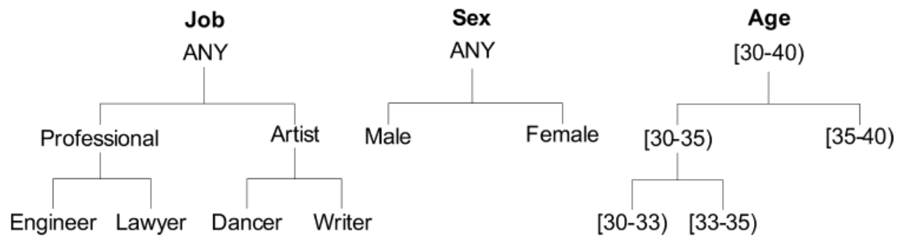


Fig. 3. Taxonomy trees for *Job*, *Sex*, *Age*.

- Full domain generalization
  - Generalize to same level in tree
- Subtree generalization
- Sibling generalization
- Cell generalization
  - Local recoding versus global recoding for above
- Multi-dimensional generalization

Generalization

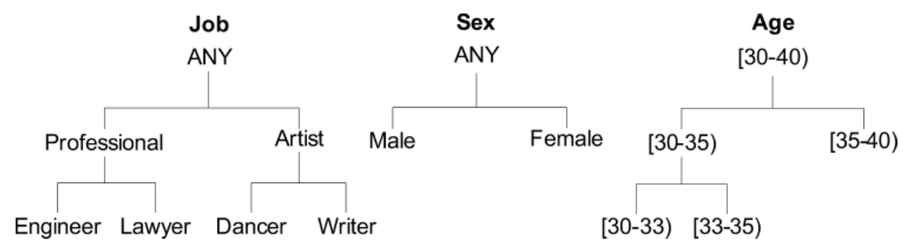


Fig. 3. Taxonomy trees for *Job*, *Sex*, *Age*.

- Full domain generalization
  - Generalize to same level in tree
- Subtree generalization
- Sibling generalization
- Cell generalization
  - Local recoding versus global recoding for above
- Multi-dimensional generalization
  - Generalize Engineer,Male -> Engineer,Any
  - Generalize Engineer,Female -> Professional,Female

Suppression

- Record suppression
- Value suppression (globally)
- Cell suppression (local value suppression)

Anatomization

Table III. Anatomy

(a) Original patient data

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

(b) Intermediate *QID*-grouped table

Age	Sex	Disease (sensitive)
[30–35)	Male	Hepatitis
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	HIV
[35–40)	Female	Flu
[35–40)	Female	Flu
[35–40)	Female	Heart
[35–40)	Female	Heart

(c) Quasi-identifier table (QIT) for release

Age	Sex	GroupID
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

(d) Sensitive table (ST) for release

GroupID	Disease (sensitive)	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

Anonymization Algortihms

Table IV. Characterization of Anonymization Algorithms

Algorithm	Operation	Metric	Optimality
Record Linkage			
Binary Search [Samarati 2001]	FG,RS	*MD*	optimal
MinGen [Sweeney 2002b]	FG,RS	*MD*	optimal
Incognito [LeFevre et al. 2005]	FG,RS	*MD*	optimal
K-Optimize [Bayardo and Agrawal 2005]	SG,RS	*DM,CM*	optimal
μ-argus [Hundepool and Willenborg 1996]	SG,CS	*MD*	minimal
Datafly [Sweeney 1998]	FG,RS	*DA*	minimal
Genetic Algorithm [Iyengar 2002]	SG,RS	*CM*	minimal
Bottom-Up Generalization [Wang et al. 2004]	SG	*ILPG*	minimal
Top-Down Specialization (TDS) [Fung et al. 2005, 2007]	SG,VS	*IGPL*	minimal
TDS for Cluster Analysis [Fung et al. 2009]	SG,VS	*IGPL*	minimal
Mondrian Multidimensional [LeFevre et al. 2006a]	MG	*DM*	minimal
Bottom-Up & Top-Down Greedy [Xu et al. 2006]	CG	*DM*	minimal
TDS2P [Wang et al. 2005; Mohammed et al. 2009]	SG	*IGPL*	minimal
Condensation [Aggarwal and Yu 2008a, 2008b]	CD	heuristics	minimal
r-Gather Clustering [Aggarwal et al. 2006]	CL	heuristics	minimal
Attribute Linkage			
Top-Down Disclosure [Wang et al. 2005, 2007]	VS	*IGPL*	minimal
Progressive Local Recoding [Wong et al. 2006]	CG	*MD*	minimal
ℓ-Diversity Incognito [Machanavajjhala et al. 2007]	FG,RS	*MD,DM*	optimal
InfoGain Mondrian [LeFevre et al. 2006b]	MG	*IG*	minimal
Anatomy [Xiao and Tao 2006a]	AM	heuristics	minimal
(k, e)-Anonymity Permutation [Zhang et al. 2007]	PM	min. error	optimal
Greedy Personalized [Xiao and Tao 2006b]	SG,CG	*ILoss*	minimal
t-Closeness Incognito [Li et al. 2007]	FG,RS	*DM*	optimal
Table Linkage			
SPALM [Nergiz et al. 2007]	FG	*DM*	optimal
MPALM [Nergiz et al. 2007]	MG	heuristics	minimal
Probabilistic Attack			
Cross-Training Round Sanitization [Chawla et al. 2005]	AN	statistical	N/A
ϵ-Differential Privacy Additive Noise [Dwork 2006]	AN	statistical	N/A
$\alpha\beta$ Algorithm [Rastogi et al. 2007]	AN,SP	statistical	N/A

FG = Full-domain Generalization, SG = Subtree Generalization, CG = Cell Generalization, MG = Multidimensional Generalization, RS = Record Suppression, VS =Value Suppression, CS = Cell Suppression, AM = Anatomization, PM = Permutation, AN = Additive Noise, SP = Sampling, CD = Condensation, CL=Clustering