K-Anonymity & Algorithms

Guillaume Raschia — Polytech Nantes; Université de Nantes

Last update: April 14, 2021

Contents

Introduction

k-Anonymity

- Privacy Model
- Generalization Mechanism
- Formal Risk Assessment

Utility Metrics

k-Anonymization Algorithms

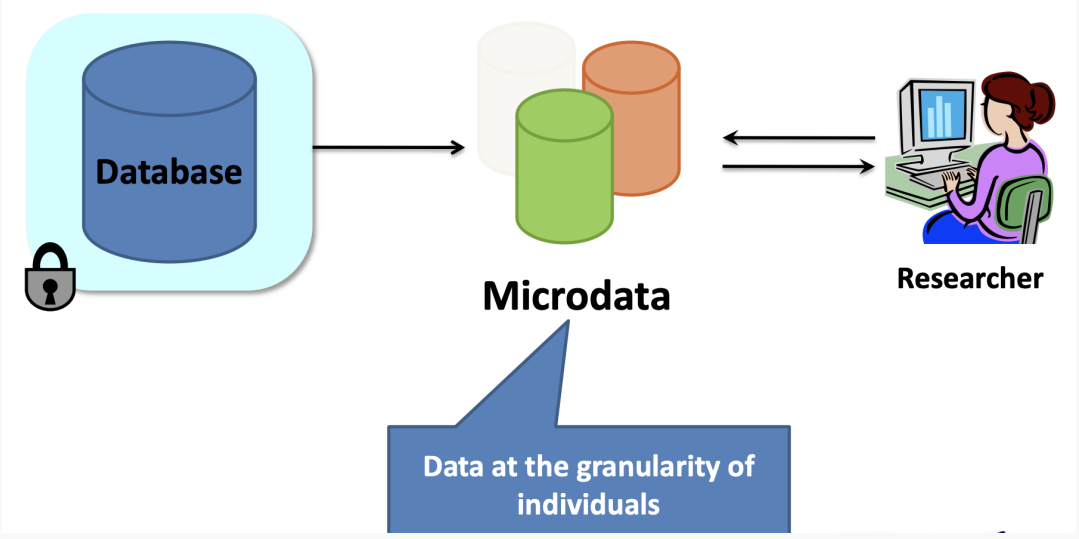
Limitations

[Source : A. Machanavajjhala - Duke Univ., 2012]

Introduction

Offline Data Publishing

Privacy-Preserving Data Publishing (PPDP)
the Non-Interactive (aka Offline) Scenario



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
248-22-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer
231-43-4582	13068	33	Chinese	Cancer

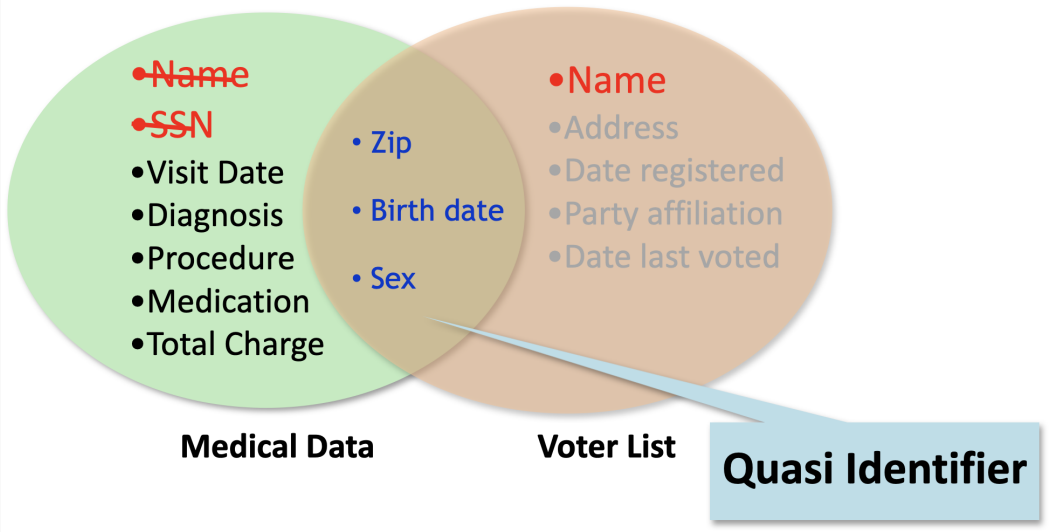
Removing SSN...

De-Identification

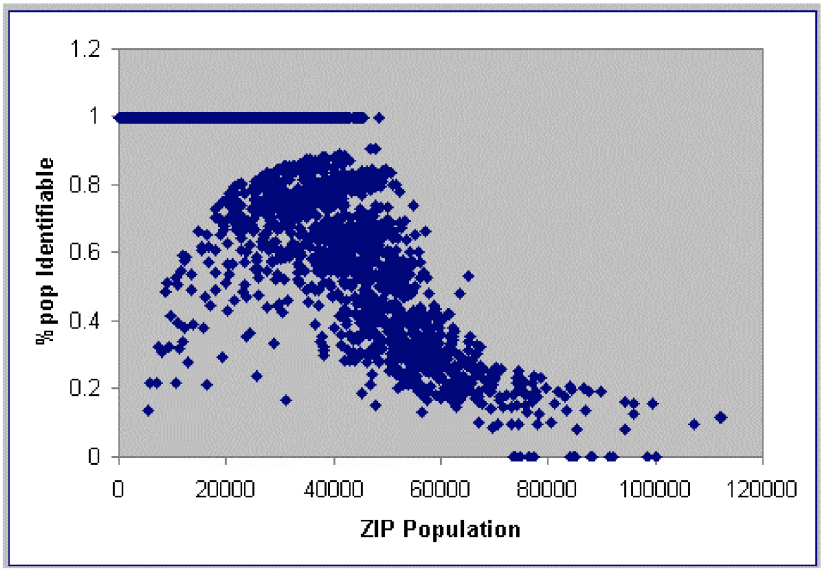
Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer
13068	33	Chinese	Cancer

The Massachussets Governor

Privacy Breach [L. Sweeney, IJUFKS 2002]



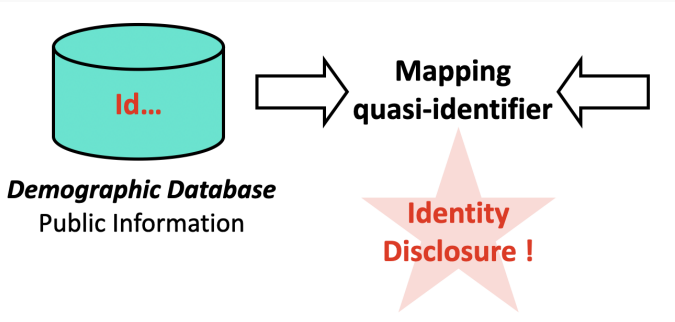
Simple Demographics Often Identify People Uniquely



https://dataprivacylab.org/projects/identifiability/paper1.pdf [L Sweeney, 2000]

Linkage Attack

- Identity Disclosure Threat



Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
...	...	...	...

Other Well-Known Linkage Attacks

- AOL (live dataset `http://search-id.com`)
  TNYT. A Face Is Exposed for AOL Searcher No. 4417749, 2006
- Netflix
  A. Narayanan,V. Shmatikov. How To Break Anonymity of the Netflix Prize Dataset, 2006
- 4 GSM points with timestamps = 95% unique
  Y.-A. de Montjoye et al., Unique in the Crowd: The privacy bounds of human mobility, Nature, 2013
- 2 French children DoB is a unique footprint of mother id
- etc

k-Anonymity

k-Anonymity

Main Idea

Generalize, modify, or distort quasi-identifier values so that any individual is indistinguishable in a group of **at least** k

- In SQL, table T is k-anonymous if each value from

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

  is $\geq k$
- Parameter k indicates the "degree" of anonymity

Generalization (Coarsening)

ORIGINAL MICRODATA

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
13068	33	Chinese	Cancer

▷

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Viral
130**	<30	Any	Viral
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

Equivalence Class: group of k-anonymous records that share the same quasi-identifier value

Alternative Transformations

- Suppression
  - Remove non-anonymous values or tuples
  - Can be combined with generalization btw of a suppression threshold
  - Handle outliers
- Clustering
- Microaggregation

Generalization-Suppression Example

ORIGINAL MICRODATA

	ID	QID			SA
num	Name	Marital Stat	Age	Zip	Crime
01	Amy	Maried	35	32042	Murder
02	Bob	Single	20	32021	Theft
03	Carl	Widowed	54	31024	Trafic
04	David	Separated	22	32026	Assault
05	Ethan	Maried	31	32025	Piracy

2-ANONYMOUS RELEASE WITH SUPPRESSION

		QID			SA
num	group	Marital Stat	Age	Zip	Crime
01	1	Maried	30s	320**	Murder
05		Maried	30s	320**	Piracy
02	2	Not Maried	20s	3202*	Theft
04		Not Maried	20s	3202*	Assault
03	3	*	*	*	*

Quasi-Identifier

Attributes of a given dataset are either:

- Direct Identifier: removed
- Quasi-Identifier (QID): transformed
- Sensitive: preserved

How to set up QID?

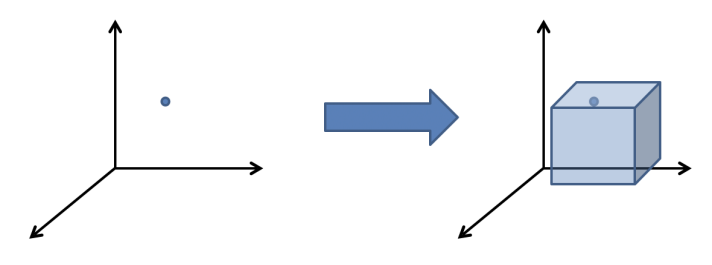
- QID is a combination of attributes (that an adversary may know) that uniquely identify a large fraction of the population
- There can be many sets of QID: if $Q = \{A, B, C\}$ is a quasi-identifier, then $Q \cup \{D\}$ is also a quasi-identifier
- Need to guarantee k-anonymity against the largest QID

Generalization

- Coarsen (or suppress) an attribute to a more general value
- '→' means one generalization step

Numeric Values

- Mask low significant bits: 12345 → 1234* → 123**
- Ranges : 23 → [20-25]; or even (30.5N20.3E) → box(30N-31N,20E-22E)

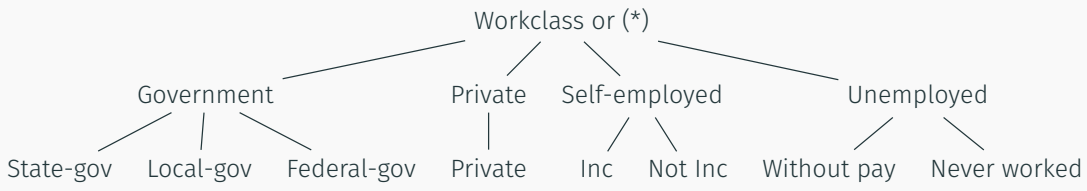


Generalization (cont'd)

- Coarsen (or suppress) an attribute to a more general value
- '→' means one generalization step

Categorical Values

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



Generalization Models

Global Recoding

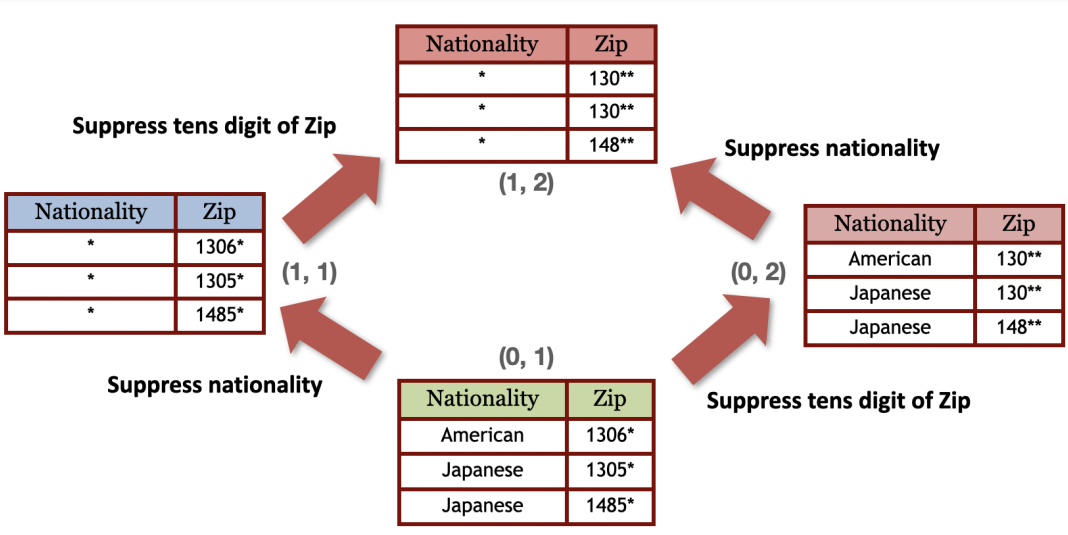
- There is one single generalized value for each value in an attribute
  - Every occurrence of 12345 is replaced with 1234* in the database
- Full domain generalization is the main instance of global recoding
  - Generalize all the values in an attribute to the same "level"
  - Answering queries on such datasets is easier

Local Recoding

- Records may be independently generalized one with each other

Generalization Lattice

- Generalization step $D \rightarrow D'$ on one attribute
- Generalization vector : each value is the level of generalization



Risk Assessment

- Denote by f_i the frequency of a QID value i in the (anonymized) microdata D
- Risk for individuals that generalize to that QID value i is then $1/f_i$

Prosecutor Risk

- Target a specific individual
- Based on microdata uniqueness
- Is the worst case re-identification scenario for the individuals :

$$\frac{1}{\min_i f_i}$$

From Prosecutor Risk to k Parameter

k-Anonymity Guarantee

- Joining k-anonymous D to a demographic dataset using quasi-identifier results in at least k records per quasi-identifier value

How to retrieve k from microdata?

- Overall k value of D is the smallest size of all the equivalence classes :

$$k = \min_i f_i$$

- Prosecutor risk is then $1/k$

k-Map and Alternative Risk Scenarios

POPULATION U

num	Zip	Age	Nationality
01	13053	28	Russian
02	13068	29	American
03	13068	21	Japanese
04	13053	23	American
05	14853	50	Indian
06	14853	55	Russian
07	14850	47	American
08	14850	59	American
09	13053	31	American
10	13053	37	Indian
11	13068	36	Japanese
12	13068	32	American
13	13068	33	Chinese

$\triangleright$

4-ANONYMOUS RELEASE OF U

Zip	Age	Nat.
130**	<30	Any
130**	<30	Any
130**	<30	Any
130**	<30	Any
1485*	≥40	Any
1485*	≥40	Any
1485*	≥40	Any
1485*	≥40	Any
130**	[30,40)	Any
130**	[30,40)	Any
130**	[30,40)	Any
130**	[30,40)	Any
130**	[30,40)	Any

k-Map

4-ANONYMOUS RELEASE OF U

num	Zip	Age	Nat.
01	130**	<30	Any
02	130**	<30	Any
03	130**	<30	Any
04	130**	<30	Any
05	1485*	≥40	Any
06	1485*	≥40	Any
07	1485*	≥40	Any
08	1485*	≥40	Any
09	130**	[30,40)	Any
10	130**	[30,40)	Any
11	130**	[30,40)	Any
12	130**	[30,40)	Any
13	130**	[30,40)	Any

$\triangleright$

4-ANONYMOUS RELEASE OF SAMPLE $D \subseteq U$

num	Zip	Age	Nat.	Disease
01	130**	<30	Any	Heart
06	1485*	≥40	Any	Heart
08	1485*	≥40	Any	Viral
09	130**	[30,40)	Any	Cancer
10	130**	[30,40)	Any	Cancer
11	130**	[30,40)	Any	Cancer
12	130**	[30,40)	Any	Cancer
13	130**	[30,40)	Any	Cancer

Alternative Risk Scenarios

Let F_i be the frequency of QID i in the population U;

Re-identification risk in the k-map setting

1. Journalist : target an arbitrary individual

$$\frac{1}{\min_i F_i} \quad \text{where } i \text{ are QID's from } D$$

   - The hard part : What is U? How to compute F_i?
   - Use estimators $\hat{F}_i$

2. Marketer : expect a high average risk

$$\frac{1}{|D|} \cdot \sum_i \frac{f_i}{F_i}$$

Utility Metrics

Quantifying Error

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

Utility Metrics

- Average size of equivalence classes : $C_{\text{avg}}(D) = \frac{|D|}{\#\text{groups} \cdot k}$
- Discernibility metric

$$C_{\text{dsc}}(D) = \sum_{f_i \geq k} f_i^2 + \sum_{f_i < k} f_i \cdot |D|$$

  - Assign a penalty to each tuple
  - Penalty depends on how many other tuples are indistinguishable from it

Example

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Viral
130**	<30	Any	Viral
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

- $C_{\text{avg}}(D) = \frac{13}{3 \times 4} = 1.08$
  - ideal value is 1.0
  - greater values mean less utility
- $C_{\text{dsc}}(D) = 2 \times 4^2 + 5^2 = 57$
  - expect a value close to $k \cdot |D|$ i.e., $4 \times 13 = 52$

Utility Metrics (cont'd)

- Number of steps in the generalization lattice
  - Precision metric : $C_{\text{prc}}(D) = 1 - \frac{1}{\#\text{cells}} \cdot \sum_{\text{cells}} \frac{\text{level}}{\text{height}}$
- Shape of equivalence classes
  - Numeric range or size of the domain
  - Penalty for each tuple $= 1 - 1/\#\text{values that can generalize to that tuple}$
  - E.g., Penalty (14850, 47) $= 1 - 1/1 = 0$
  - Penalty (1485*, [40-50)) $= 1 - 1/(10 \times 10) = 0.99$
  - Generalized Information Loss :

$$C_{\text{gil}}(D) = \frac{1}{\#\text{cells}} \cdot \sum_{\text{cells}} \frac{|\text{cell}|}{|\text{dom}|}$$

Example

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Viral
130**	<30	Any	Viral
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

- $C_{\text{prc}}(D) = 1 - \frac{1}{13\times3} \cdot (\frac{9\times2}{5_{\text{Z}}} + \frac{4}{5_{\text{Z}}} + \frac{13}{3_{\text{A}}} + \frac{13}{1_{\text{N}}}) = 0.44$
  - depends on the height of each DGH : 5_{Z}, 3_{A}, 1_{N}
  - range in $[0.0, 1.0]$ from low to high utility
- $C_{\text{gil}}(D) = \frac{1}{39} \cdot (\frac{9\times100}{100\,000_{\text{Z}}} + \frac{4\times10}{100\,000_{\text{Z}}} + \frac{4\times30}{100_{\text{A}}} + \frac{4\times60}{100_{\text{A}}} + \frac{5\times10}{100_{\text{A}}} + \frac{13\times10}{10_{\text{N}}}) = 0.438$

Utility Metrics (cont'd)

Classification Metrics

- Assign a penalty to each tuple t :
  - If t's sensitive value is the majority value in the group: penalty is null
  - Otherwise, penalty is the size of equivalence class f_i

QID Distribution Metrics

- Assessing the change in the distribution of the underlying data
- Kullback-Leibler Divergence

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 $\hat{F}$

$\hat{F}$(14850, 47, American) = 1/12 i.e., the fraction of tuples in the database with

- Zip = 14850 AND Age = 47 AND Nationality = American

KL-Divergence (cont'd)

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

$$\hat{F}_{k\text{anon}}(14850, 47, \text{American}) = \frac{1}{|D|} \cdot \sum_{\text{group } i} P[(14850, 47, \text{American}) \in i] \times f_i$$

Example

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Viral
130**	<30	Any	Viral
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

- $\hat{F}_{k\text{anon}}(14850, 47, \text{American}) = \frac{1}{13} \cdot P[(14850, 47, \text{American}) \in c_2] \times 4 = \frac{1}{13} \cdot \frac{4}{10 \times 60 \times 10} = 5 \cdot 10^{-5}$

KL-Divergence (cont'd)

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

$$\sum_x p(x) \cdot \log \frac{p(x)}{p_{k\text{anon}}(x)}$$

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

k-Anonymization Algorithms

k-Anonymization Problem

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

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

NP-Hard Problem

[Meyerson and Williams, PODS 2004]

- Reduction from the k-dimensional matching problem
- There is a $\log k$ approximation algorithm for some utility metrics

Algorithms

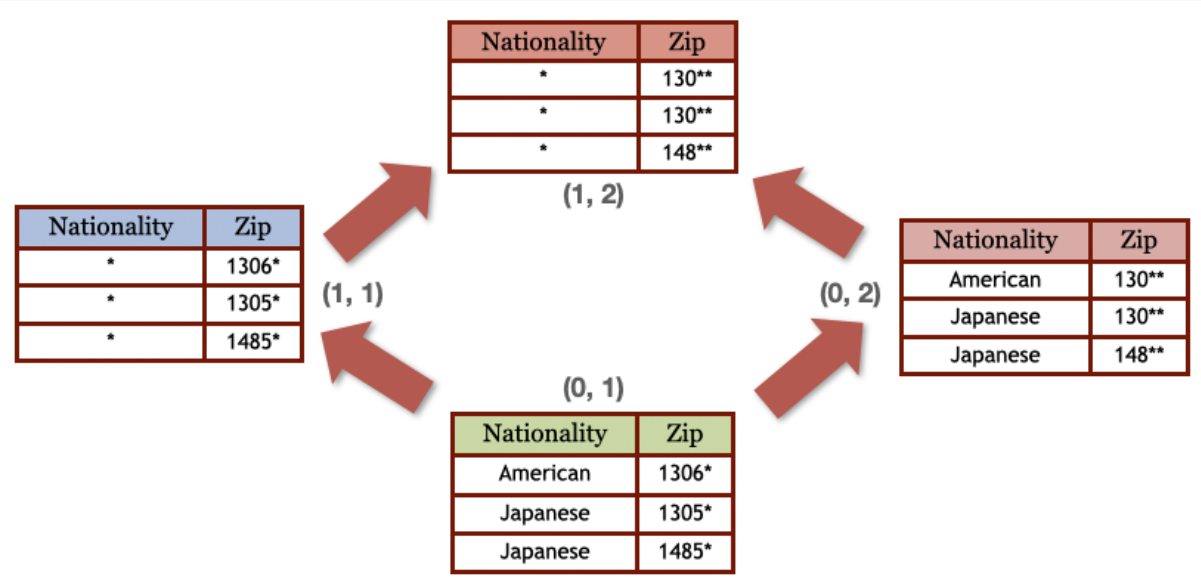
Optimal

- allMin [Samarati, 2001]
- minGen [Sweeney, 2002]
- Incognito [LeFevre et al, 2005]
- k-OPTIMIZE [Bayardo et al, 2005]
- Flash [Kohlmayer et al, 2012]
- etc

Approximate

- Datafly [Sweeney, 1997]
- Mondrian [LeFevre et al, 2006]
- Hilbert [Ghinita et al, 2007]
- BangA [Anjum et al, 2017]
- etc

Generalization Lattice : Gentle Reminder



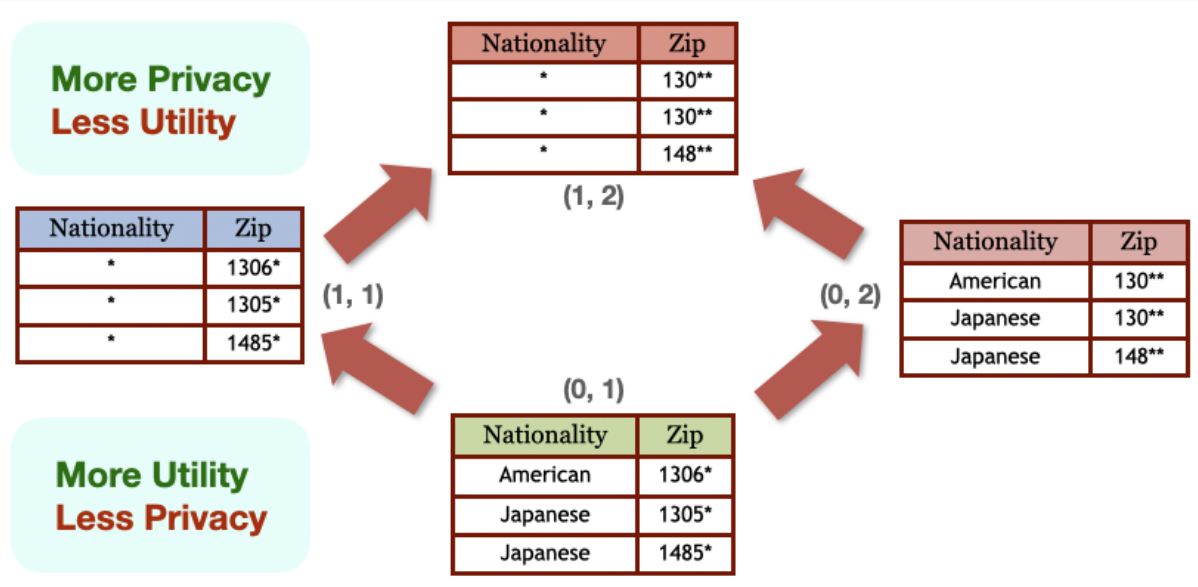
A Very First Naive Algorithm

minGen

1. Enumerate each generalization (randomly traverse the all lattice)
2. Check for k-anonymity: discard non-anonymous generalizations
3. Evaluate utility and rank candidate releases
4. Pick one with the highest utility

- minGen stands for k-minimal Generalization

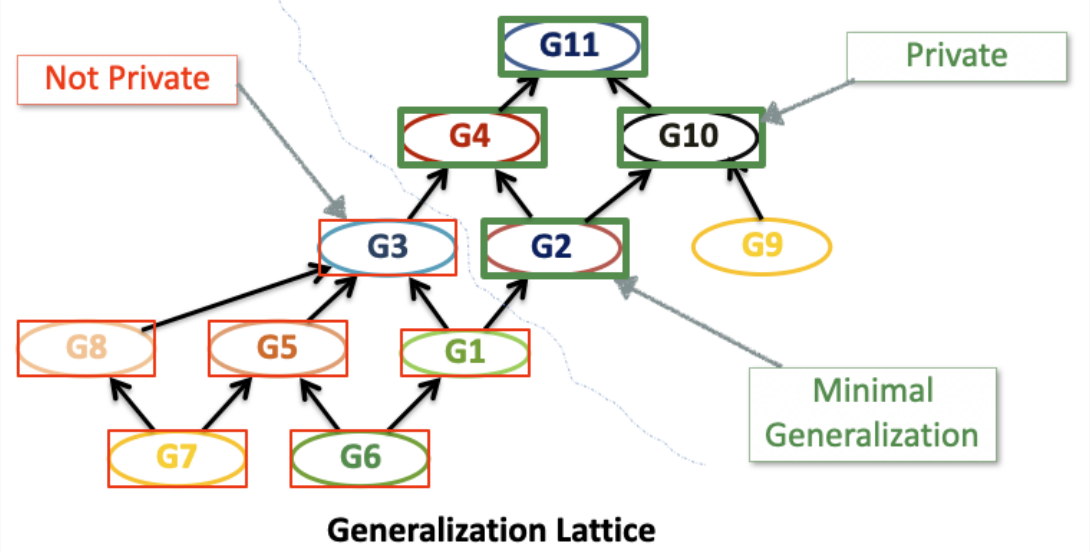
Monotonicity



Monotonicity (cont'd)

- In a single generalization step $D \rightarrow D'$, 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 lower utility than D
  - Intuitively true: more information is hidden in D'
  - Can be formally shown for all the utility metrics discussed
- What is the impact of suppression on monotonicity?

Pruning Using Monotonicity

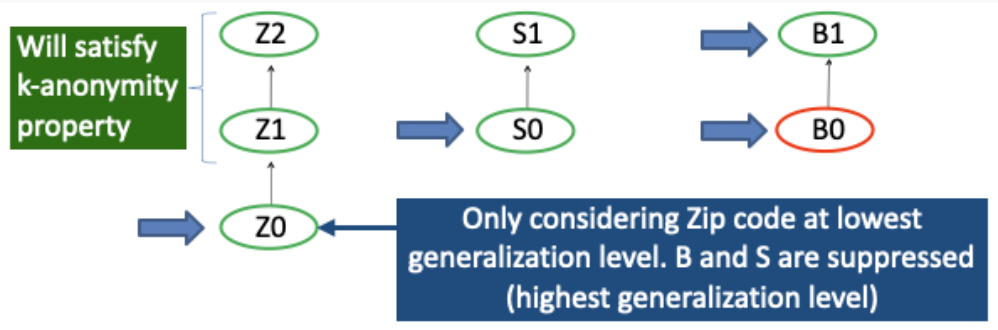


Basic Incognito Algorithm

Zip-Sex-Birthdate Microdata

Step 1

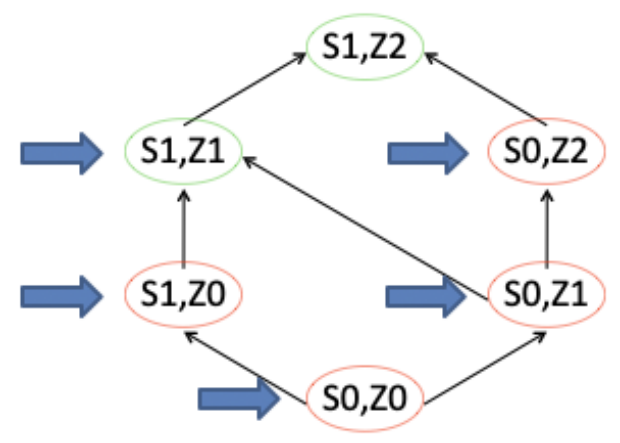
- Start with 1 dimensional quasi-identifier
- Traverse bottom-up to check when k-anonymity is satisfied



Basic Incognito Algorithm (cont'd)

Step 2

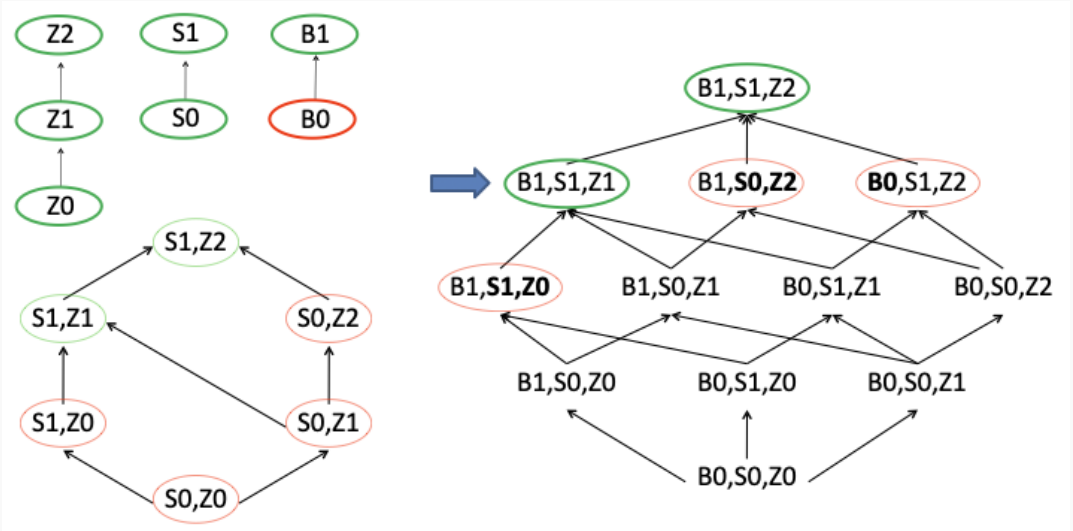
Move to 2 dimensional marginals



Basic Incognito Algorithm (cont'd)

Step 3

3 dimensional quasi-identifier



Summary of Incognito Algorithm

Problem

Amongst all tables that satisfy k-anonymity, find the one that has maximum 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)

Datafly Algorithm: Example from Sweeney, 2002

Race	BirthDate	Gender	ZIP	#occurs	
black	9/20/65	male	02141	1	t1
black	2/14/65	male	02141	1	t2
black	10/23/65	female	02138	1	t3
black	8/24/65	female	02138	1	t4
black	11/7/64	female	02138	1	t5
black	12/1/64	female	02138	1	t6
white	10/23/64	male	02138	1	t7
white	3/15/65	female	02139	1	t8
white	8/13/64	male	02139	1	t9
white	5/5/64	male	02139	1	t10
white	2/13/67	male	02138	1	t11
white	3/21/67	male	02138	1	t12
2	12	2	3		

A

Race	BirthDate	Gender	ZIP	#occurs	
black	1965	male	02141	2	t1,t2
black	1965	female	02138	2	t3, t4
black	1964	female	02138	2	t5, t6
white	1964	male	02138	1	t7
white	1965	female	02139	1	t8
white	1964	male	02139	2	t9, t10
white	1967	male	02138	2	t11, t12
2	3	2	3		

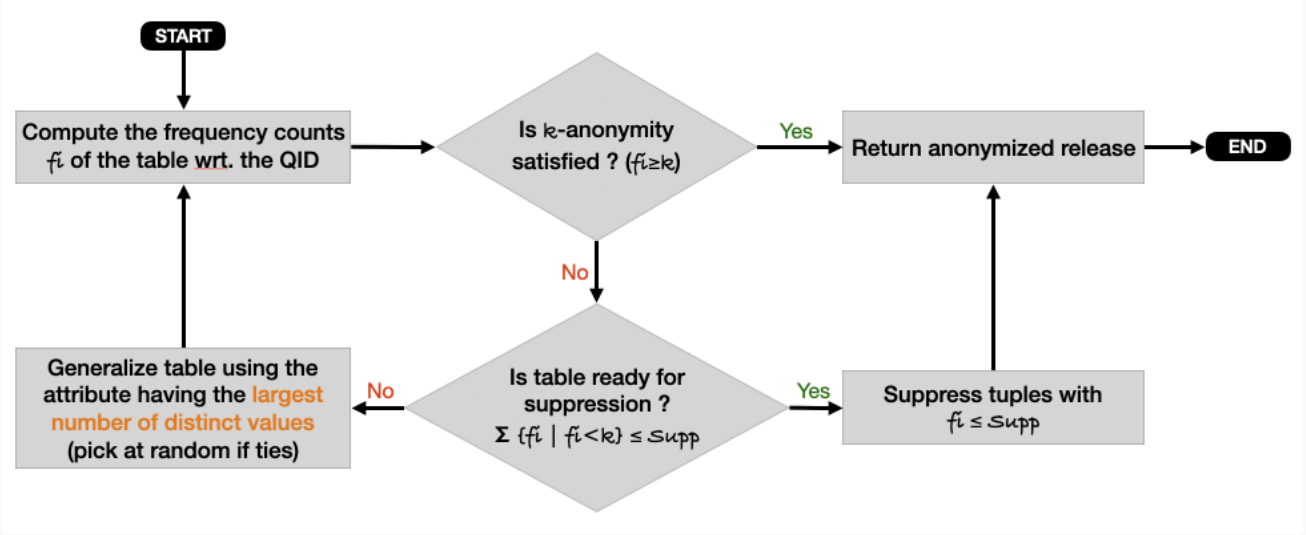
B

Figure 9 Intermediate stages of the core Datafly algorithm

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

Figure 10 Table MGT resulting from Datafly, k=2, QI={*Race*, *Birthdate*, *Gender*, *ZIP*}

Datafly Algorithm



Limitations

Attack 1: Homogeneity

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Viral
130**	<30	Any	Viral
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

Name	Zip	Age	Nat.
Bob	13053	35	French

- Bob has cancer

Attack 2: Background Knowledge

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Flu
130**	<30	Any	Flu
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

- Japanese have a very low incidence of Heart disease
- Umeko has flu

k-Anonymity Limitations

Attribute Disclosure

- Homogeneity Attack
- Background Knowledge Attack
- "Gaydar"-like inference of sensitive values in Online Social Networks
  Zheleva and Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles, WWW 2009

✓ ℓ-diversity, t-closeness, etc

k-Anonymity Limitations (cont'd)

Multiple Releases

- Linkage Attack using successive k-anonymous releases

✓ m-uniqueness, τ-safety

Syntactic/Deterministic Approach Only

- Attacker gains knowledge about individuals thx to the k-anonymous release

✓ Differential Privacy comes to the rescue