Data Privacy
Hiding Data from the Database User I

Erman Ayday

Some slides from
Vitaly Shmatikov – UT Austin
Murat Kantarcioglu – UT Dallas

Databases

- Many databases contain sensitive (personal) data
  - Hospital records, internet search information, the set of friends on different social sites, etc.
- It is a common scenario that the release of a function/ statistic on such data is socially beneficial
  - Used for apportioning resources, evaluating medical therapies, understanding the spread of disease, improving economic utility, and informing us about ourselves as a species
  - E.g., the usage of hospital records greatly helps medical research
- Hard to publish databases in a privacy-preserving way
- Crucial to ensure that the release of a function on a database does not leak too much information about the individuals
  - Differential privacy is a quite recent notion that tries to formalize this requirement

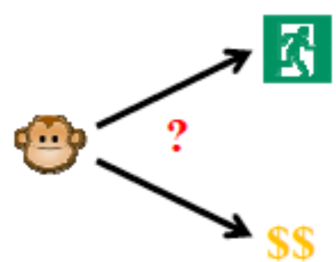
Natural Sources of Big Data

Social networks & media

Recommender systems

Web tracking dbs (profiling)

Doc indexing & search



Predicting user behavior

Exposing trends

Some Examples

- Health-care datasets
  - Clinical studies, hospital discharge databases …
- Genetic datasets
  - 1000 Genome, HapMap, deCode …
- Demographic datasets
  - U.S. Census Bureau, sociology studies …
- Search logs, recommender systems, social networks, blogs …
  - AOL search data, social networks of blogging sites, Netflix movie ratings, Amazon …

What About Privacy?

- First thought: anonymize the data
- How?
- Remove "personally identifying information" (PII)
  - Name, Social Security number, phone number, email, address… what else?
  - Anything that identifies the person directly
- Is this enough?

What is Anonymous?

- One is anonymous, who can not be identified within a set of subjects
  - Anonymity set!
  - Identifying attributes are the same
  - Point of view can be local or global
  - Determined by the attacker model

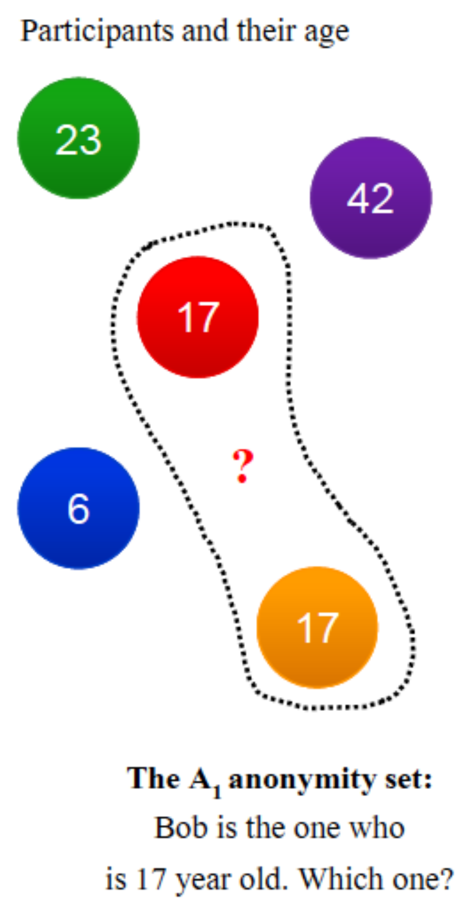
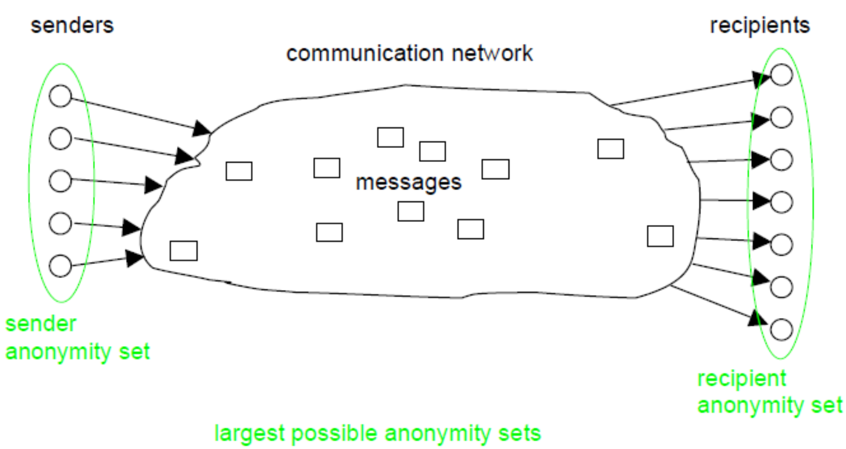
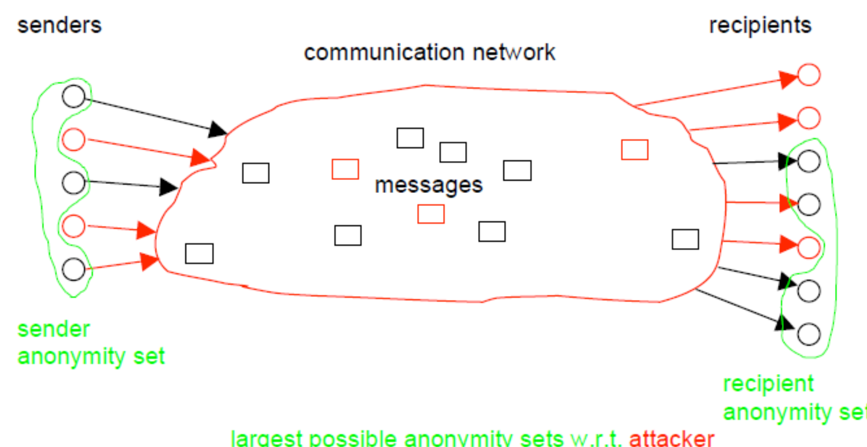


Figure: Gabor Gorgy Gulyas

Reminder - Anonymity

- Anonymity: state of being not identifiable within a set of subjects (the anonymity set)
- All other things being equal, anonymity is the stronger if
  - the respective anonymity set is larger
  - the sending or receiving of the subjects within that set is more evenly distributed





How Identifiable Are We?

Sweeney, 1990

87% of US population is identifiable by (216 million of 248 million):
`{5 digit ZIP, gender, date of birth}`

Revisiting study: 64% of US population is identifiable by:
`{ZIP-code, gender, date of birth}`

Golle, 2000

Figure: Gabor Gorgy Gulyas

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	

Figure 1: Re-identifying anonymous data by linking to external data

Public voter dataset

Privacy Mechanisms for Databases

- Non-interactive mechanisms
  - Database publishes a sanitized dataset
  - Researcher asks arbitrary queries on the sanitized dataset

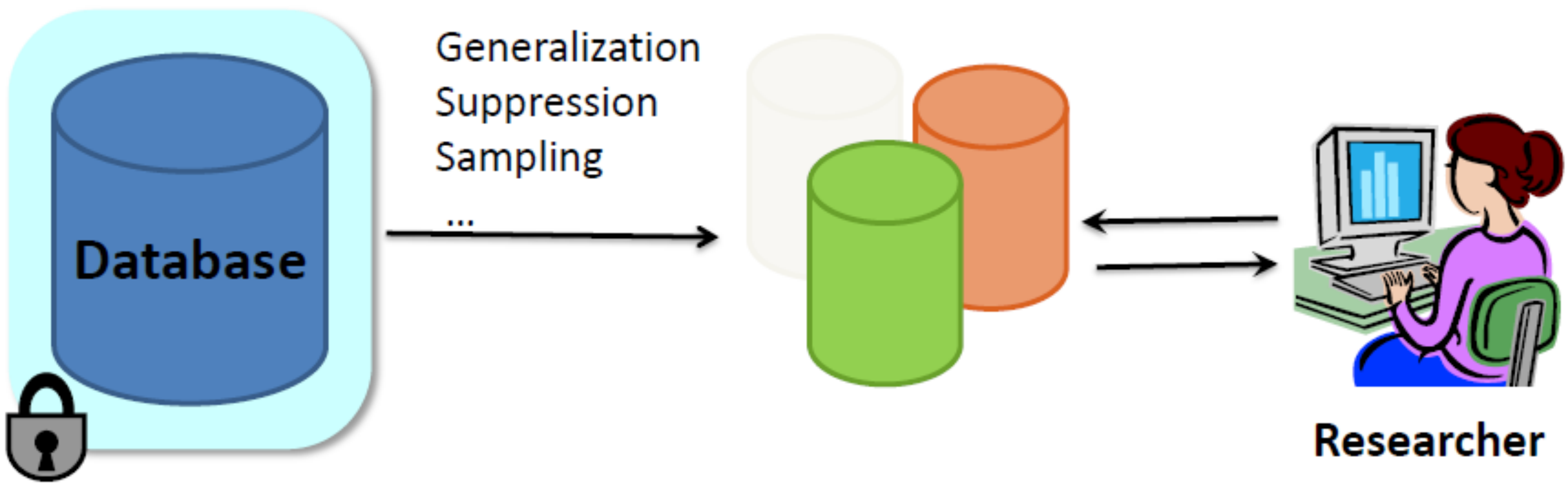


Figure: Ashwin Machanavajjhala

Privacy Mechanisms for Databases

- Interactive mechanisms
  - Researcher directly asks queries to the database
  - Database can choose to answer truthfully or answer with noise
  - Auditor may keep track of all the queries pose to the database and deny queries
- Next Class …

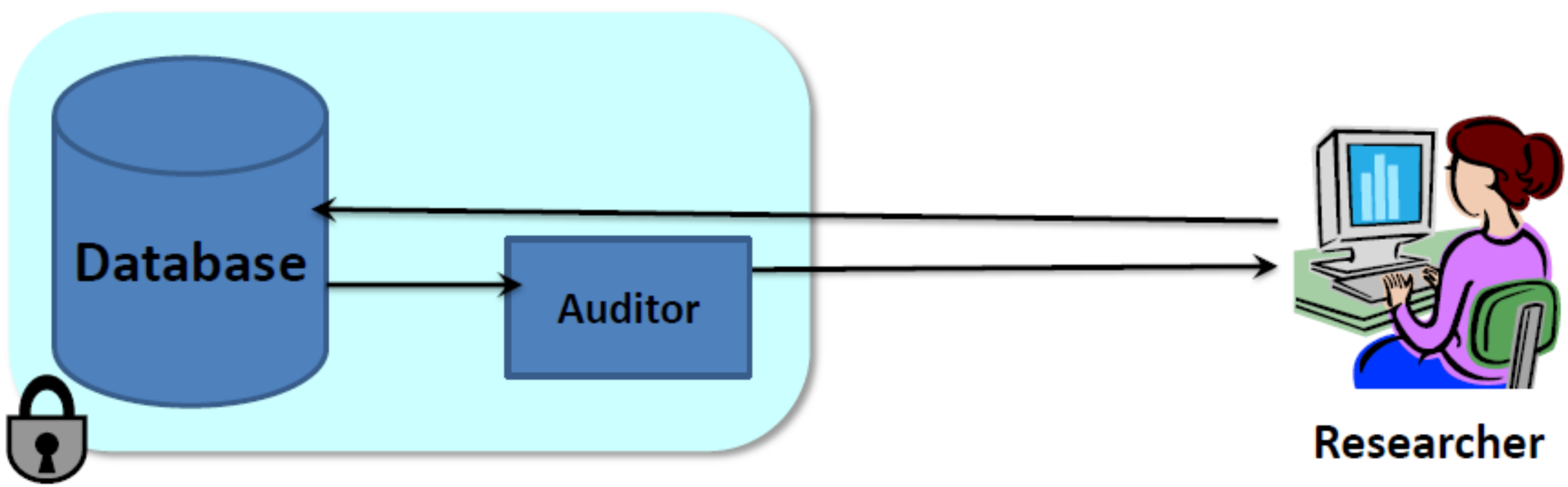


Figure: Ashwin Machanavajjhala

k-Anonymity - Overview

- The database achieves k-anonymity if for all records there are at least (k-1) other rows with the same **quasi identifier**
- Methods: supression or generalization
- Attributes can be: explicit id, quasi id, sensitive

Employee database

Name	Birth date	City
John	1980-01-31	New York
Emily	1976-06-25	Flint
Bob	1985-09-05	New York
Dave	1973-02-07	South Bend
	...	

Healthcare database

Birth date	City	Diagnosis
1985-09-05	New York	Stroke
1973-02-07	South Bend	-
1980-01-31	New York	Flu
1976-06-25	Flint	HIV
	...	

Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized dataset with other datasets

Classification of Attributes

- Sensitive attributes
  - Medical records, salaries, etc.
  - These attributes is what the researchers need, so they are always released directly

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

k-Anonymity Example

Employee database

Name	Birth date	City
John	1980-01-31	New York
Emily	1976-06-25	Flint
Bob	1985-09-05	New York
Dave	1973-02-07	South Bend

Healthcare database

Birth date	City	Diagnosis
198*	New York	Stroke
197*	South Bend	-
198*	New York	Flu
197*	Flint	HIV

?

Better: P(„John has flu")=1 → P(„John has flu")= ½

Employee database

Name	Birth date	City
John	1980-01-31	New York
Emily	1976-06-25	Flint
Bob	1985-09-05	New York
Dave	1973-02-07	South Bend

Healthcare database

Birth date	City	Diagnosis
198*	New York	Stroke
197*	[small city]	-
198*	New York	Flu
197*	[small city]	HIV

Even better: probs are now ½ for all! (2-anonymity)

Figure: Gabor Gorgy Gulyas

Example of a k-Anonymous Table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

k-Anonymity – Definition

- Each person contained in the database cannot be distinguished from at least k-1 other individuals whose information also appear in the released database

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	02141	short breath
t2	Black	1965	m	02141	chest pain
t3	Black	1964	f	02138	obesity
t4	Black	1964	f	02138	chest pain
t5	White	1964	m	02138	chest pain
t6	White	1964	m	02138	obesity
t7	White	1964	m	02138	short breath

- Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender

[1] L. Sweeney. K-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557-570, Oct. 2002

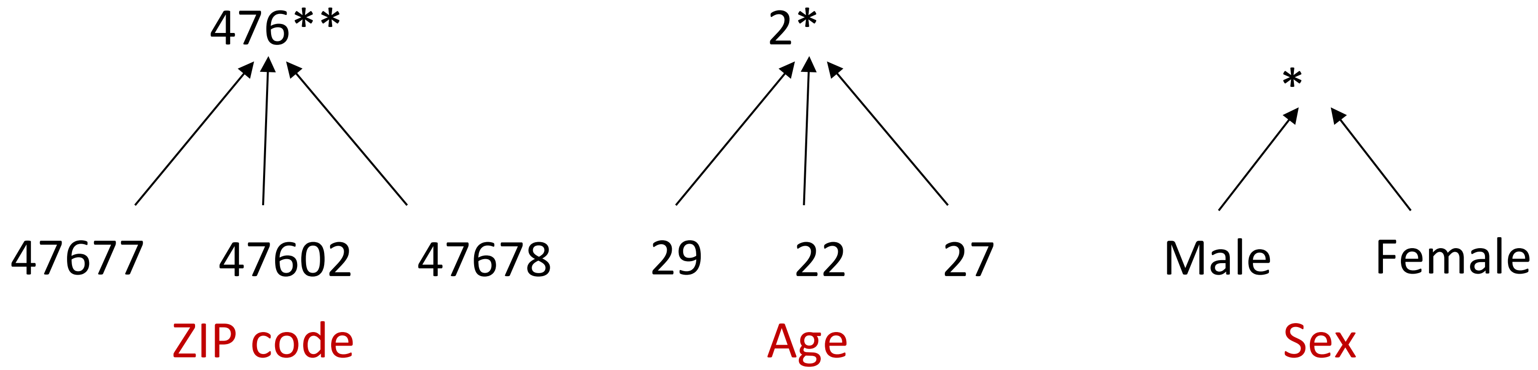
Achieving k-Anonymity

- Generalization
  - Replace specific quasi-identifiers with less specific values until get k identical values
  - Partition ordered-value domains into intervals
- Suppression
  - "Not releasing any value at all"
  - When generalization causes too much information loss
    - This is common with "outliers"
- Lots of algorithms in the literature
  - Aim to produce "useful" anonymizations

  … usually without any clear notion of utility

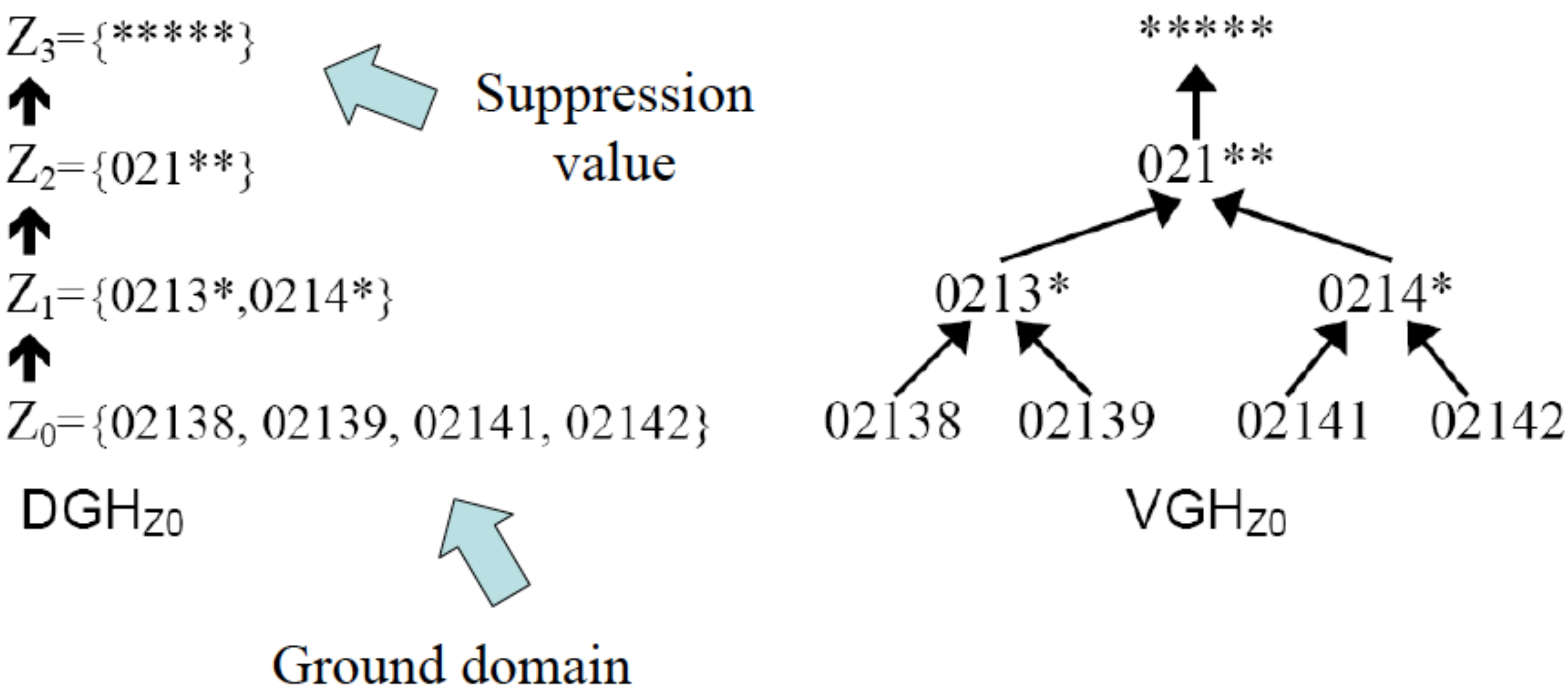
Generalization

- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- Generalization: replace quasi-identifiers with less specific, but semantically consistent values

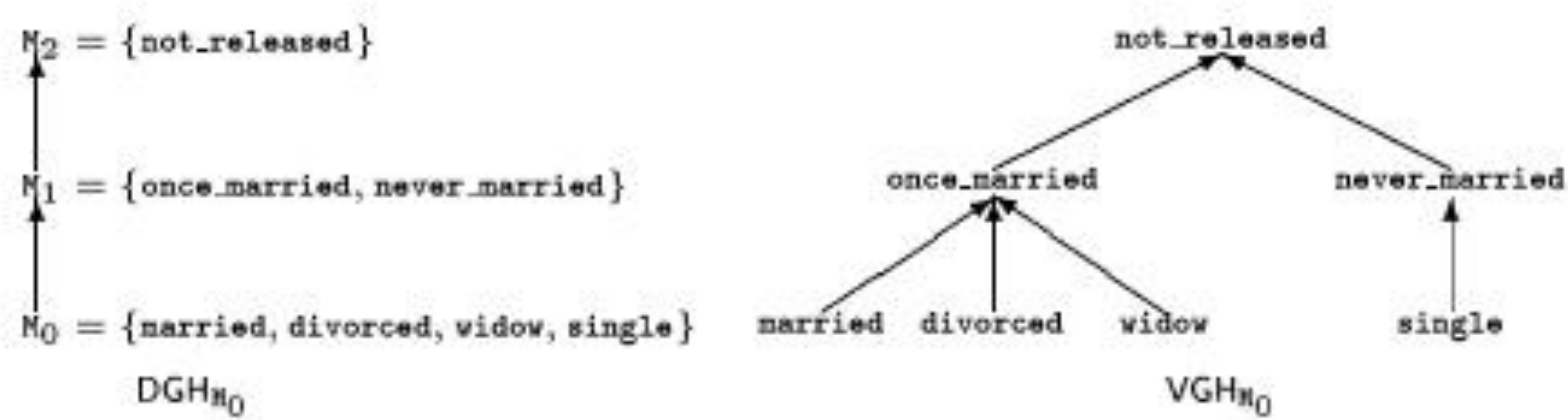


Generalization - ZIP

- ZIP attribute



Different Generalizations



Example of Generalization (1)

Released table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

External data Source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

By linking these 2 tables, you still don't learn Andre's problem

Example of Generalization (2)

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Generalized table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

!!

- Released table is 3-anonymous
- If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record

k-Anonymity via Generalization

- QI = {Race, ZIP}
- k = 2

- k-anonymous relation should have at least 2 tuples with the same values on

$$Dom(Race_i) \times Dom(ZIP_j)$$

where $Race_i$ and ZIP_j are chosen from corresponding DGHs

k-Anonymity via Generalization

Race E_0	ZIP Z_0
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

$Z_3=\{*****\}$

$\uparrow$

$Z_2=\{021**\}$

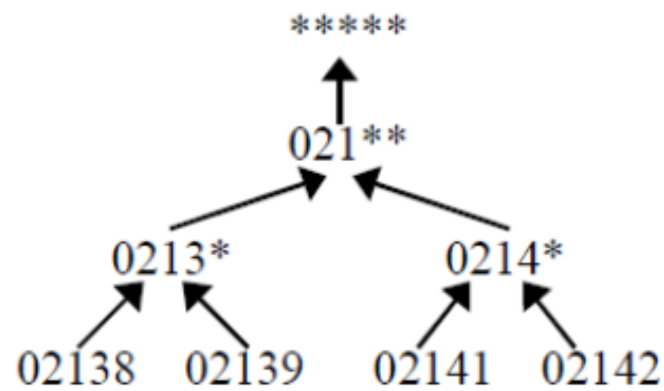
$\uparrow$

$Z_1=\{0213*, 0214*\}$

$\uparrow$

$Z_0=\{02138, 02139, 02141, 02142\}$

DGH_{Z0}



VGH_{Z0}

$Z_2=\{******\}$

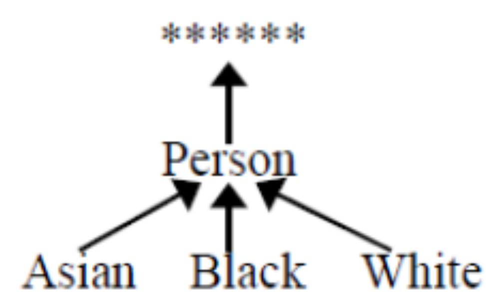
$\uparrow$

$Z_1=\{\text{Person}\}$

$\uparrow$

$Z_0=\{\text{Asian}, \text{Black}, \text{White}\}$

DGH_{E0}



VGH_{E0}

k-Anonymity via Generalization

Race E_0	ZIP Z_0
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race E_1	ZIP Z_0
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

$GT_{[1,0]}$

Race E_1	ZIP Z_1
Person	0213*
Person	0213*
Person	0214*
Person	0214*
Person	0213*
Person	0213*
Person	0214*
Person	0214*

$GT_{[1,1]}$

Race E_0	ZIP Z_2
Black	021**
Black	021**
Black	021**
Black	021**
White	021**
White	021**
White	021**
White	021**

$GT_{[0,2]}$

Race E_0	ZIP Z_1
Black	0213*
Black	0213*
Black	0214*
Black	0214*
White	0213*
White	0213*
White	0214*
White	0214*

$GT_{[0,1]}$

- The number of generalizations, enforced at the attribute level, for table T is:

$$\prod_{i=1}^{n}(|DGH_i| + 1)$$

- Total number of generalizations for PT is:

(DGH_Race+1).(DGH_ZIP + 1) = 12

Which generalization to use?

k-Minimal Generalization

- Given |R| ≥ k, there is always a trivial solution
  - Generalize all attributes to VGH root
  - Not very useful if there exists another k-anonymization with higher granularity (more specific) values

- k-minimal generalization
  - Satisfies k-anonymity
  - None of its specializations satisfies k-anonymity
  - E.g., [0,2] is not minimal, since [0,1] is k-anonymous
  - E.g., [1,0] is minimal, since [0,0] is not k-anonymous

- A table T, generalization of PT, is k-minimal if it satisfies k-anonymity and there does not exist a generalization of PT satisfying k-anonymity of which T is a generalization.

Precision Metric, Prec(.)

- Multiple k-minimal generalizations may exist
  - E.g., [1,0] and [0,1] from the example

- Precision metric indicates the generalization with minimal information loss and maximal usefulness

- Problem: how to define usefulness

Precision Metric, Prec(.)

- Precision: average height of generalized values, normalized by VGH depth per attribute per record

$$Prec(T') = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N'} \frac{h}{|DGH_{A_i}|}}{N \times N_A}$$

  - N_A : number of attributes (quasi-identifiers)
  - N: data set size (number of rows in the original table)
  - N': number of rown in the generalized table T'
  - h: generalization level of the attribute
  - |DGH(A_i) | : depth of the VGH for attribute A_i

$$Prec(T') = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N'} \frac{h}{|DGH_{A_i}|}}{N \times N_A}$$

- N = N’ if no rows of the original table are deleted/ suppressed
- When T = T’, each value is in the ground domain
  – Each h = 0, and hence Prec(T’) = 1
- When each value in T’ is the maximal element of its hierarchy
  – Each h = |DGH(A_i)|, and hence Prec(T’) = 0
- GT[1,0] and GT[0,1] each generalize values up one level
  – Since |DGH_Race| = 2 and |DGH_ZIP| = 3, Prec(GT[0,1]) > Prec(GT[1,0]).

Precision Metric, Prec(.)

- Precision depends on DGH/VGH
- Different DGHs result in different precision measurements for the same table
- Structure of DGHs might determine the generalization of choice
- DGHs should be semantically meaningful
  – I.e., created by domain experts

k-Minimal Distortion

- Most precise release that adheres to k-anonymity
- Precision measured by *Prec(.)*
- Any k-minimal distortion is a k-minimal generalization

- In the example, only [0,1] is a k-minimal distortion
  - [0,0] is not k-anonymous
  - [1,0] and others are less precise

Complexity

- Given some data set *R* and a QI *Q*, does *R* satisfy k- anonymity over *Q*?
  - Easy to tell in polynomial time
- Finding an *optimal* anonymization is not easy
  - NP-hard: reduction from k-dimensional perfect matching
- Heuristic solutions exist
  - DataFly, Incognito, Mondrian, etc.

MinGen Algorithm

- Exhaustive search
- Creates all possible generalizations of a dataset
- Picks the one that satisfies k-anonymity with minimal distortion
- Lack efficiency, especially for high number of quasi-identifiers

DataFly Algorithm

- Step 1: constructs a list *freq*
  - A frequency list containing distinct sequences of values from a private table T, along with the number of occurrences of each sequence
- Step 2: the attribute having the highest number of distinct values in *freq* is generalized
  - Continue until there remains k or fewer tuples having distinct sequences in freq
- Step 3: suppress (i.e., remove) any sequences of *freq* occurring less than k times
- Can over-distort the data when providing k-anonymity

Incognito

- Domain generalization hierarchies of the individual attributes are combined to form a multi-attribute generalization lattice
- Begins by checking single-attribute subsets of the quasi-identifiers
- Iterates, checking k-anonymity with respect to increasingly large subsets

k-Anonymity - Limitations

- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info $\Rightarrow$ k-anonymized datasets are useless

Things to be Careful About

- Unsorted Matching Attack
- Complementary Release Attack
- Linking Independent Releases

Unsorted Matching Attack

- Problem: records appear in the same order in the released table as in the original table
- Solution: randomize order before releasing

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Complementary Release Attack

- Different releases of the same private table can be linked together to compromise k-anonymity

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

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	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Use the better background knowledge attack

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Japanese Umeko has viral infection

Neighbor Bob has cancer

Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Umeko	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	⩾40	Flu
4790*	⩾40	Heart Disease
4790*	⩾40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

k-Anonymity Discussion

- These attacks show that in addition to k-anonymity, the sanitized table should also ensure diversity
- All tuples that share the same values of their quasi-identifiers should have diverse values for their sensitive attributes
- l-diversity

l-Diversity

- An equivalence class is said to have l-diversity if there are at least l well-represented values for the sensitive attribute
- A table is said to have l-diversity if every equivalence class of the table has l-diversity.

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

A 3-diverse hospital records dataset

[1] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, Mar. 2007

l-Diversity Variations

- Distinct l-Diversity
- Entropy l-Diversity
- Recursive (c,l)-Diversity

Distinct l-Diversity

- Each equivalence class has at least l well-represented sensitive values
- Doesn't prevent probabilistic inference attacks

…	Disease
	…
	HIV
	HIV
	…
	HIV
	pneumonia
	bronchitis
	…

10 records

8 records have HIV

2 records have other values

Entropy l-Diversity

- In each equivalence class, different sensitive values must be distributed evenly
- The entropy of the distribution of sensitive values in each equivalence class is at least log(l)
- Entropy of an equivalence class:

$$Entropy(E) = -\sum_{s \in S} p(E,s) \log p(E,s)$$

  - p(E,s): fraction of records in E that have sensitive value s.
- May be too restrictive
  - The entropy of the entire table may be low if a few values are very common

Recursive (c,l)-Diversity

- $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$
  - r_i is the frequency of the i^{th} most frequent value
  - m: number of distinct sensitive attributes in an equivalence class
  - Should hold for all equivalence classes
- Intuition: the most frequent value does not appear too frequently
  - And the less frequent values do not appear too rarely.

l-Diversity Limitations

Original dataset

…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Flu

99% have cancer

Anonymization A

Q1	Flu
Q1	Flu
Q1	Cancer
Q1	Flu
Q1	Cancer
Q1	Cancer
Q2	Cancer

Anonymization B

Q1	Flu
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q2	Cancer
Q2	Flu

99% cancer ⇒ quasi-identifier group is not "diverse"
…yet anonymized database does not leak anything

50% cancer ⇒ quasi-identifier group is "diverse"
This leaks a ton of information

l-Diversity Limitations

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
  - Very different degrees of sensitivity!
- l-diversity is unnecessary
  - 2-diversity is unnecessary for an equivalence class that contains only HIV- records
- l-diversity is difficult to achieve
  - Suppose there are 10000 records in total
  - To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes

Skewness Attack

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records
- Diverse, but potentially violates privacy!
- l-diversity does not differentiate:
- Equivalence class 1: 49 HIV+ and 1 HIV-
- Equivalence class 2: 1 HIV+ and 49 HIV-

l-diversity does not consider overall distribution of sensitive values!

Similarity Attack

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	⩾40	50K	Gastritis
4790*	⩾40	100K	Flu
4790*	⩾40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider semantics of sensitive values!

l-Diversity Discussion

- k-anonymity prevents identity disclosure but not attribute disclosure
- To solve that problem l-diversity requires that each eq. class has at least l values for each sensitive attribute
- But l-diversity has some limitations
- **t-closeness** requires that the distribution of a sensitive attribute in any eq. class is close to the distribution of a sensitive attribute in the overall table

t-Closeness

- An eq. class has t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t
- A table has t-closeness if all equivalence classes have t-closeness
- To measure the distance between two distributions: “earth mover distance”
  - Minimal amount of work needed to transform one distribution to another by moving distribution mass between each other

N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. IEEE 23rd Intl Conf. on Data Engineering (ICDE), 2007

t-Closeness

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database

Similarity Attack Example

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Anonymous, “t-Close” Dataset

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is k-anonymous,
l-diverse and t-close…

…so secure, right?

What Does Attacker Know?

Bob is Caucasian and I heard he was admitted to hospital with flu…

This is against the rules!
"flu" is not a quasi-identifier

Yes… and this is yet another problem with k-anonymity

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
		HIV+	Shingles
Caucas	787XX	HIV-	Acne
		HIV-	Shingles
Caucas	787XX	HIV-	Acne

Structural De-anonymization in Social Networks

- Privacy Properties
  - Social network = nodes, edges (relationships between nodes), and information associated with each node and each edge
  - Information about nodes obviously wants to satisfy a level of privacy
  - Most social networks make relationships between nodes public by default (few users change)

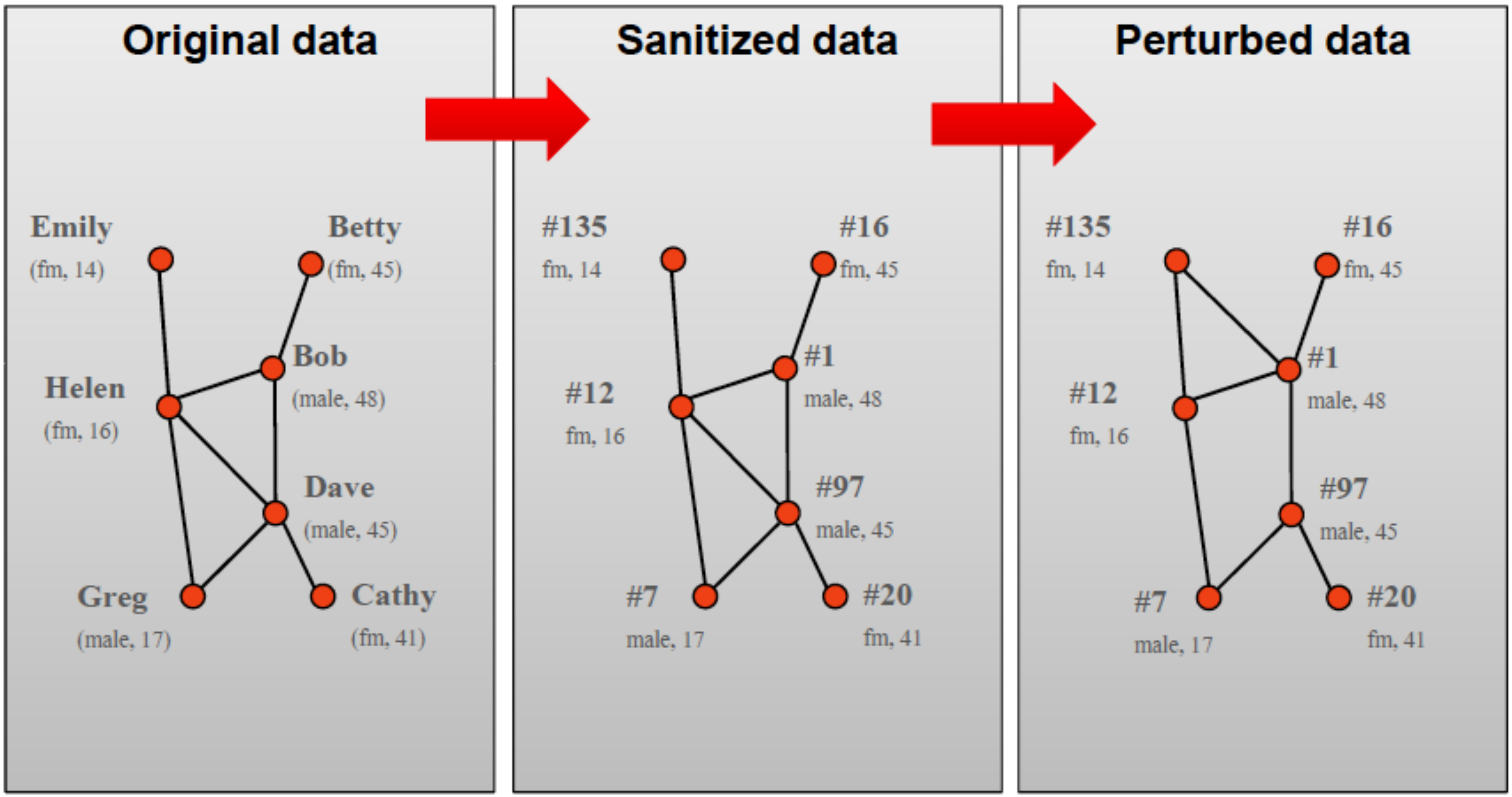
Model – Social Network

- Let us define a social network *S* consists of
  1. A directed graph G = (V,E)
  2. A set of attributes X for each node in V and a set of attributes Y for each edge in E

Attributes for nodes: (i.e. name, telephone #)

Attributes for edges: (i.e. type of relationship)

Graph Sanitization and Perturbation



Attacker Model

- Assume an attacker has access to an anonymized, sanitized, target network S_{SAN} and also access to a different network S_{AUX} whose members partially overlap with S_{SAN}
- This is a very real and plausible assumption
- Facebook -> Myspace or Twitter -> Flickr
- Even with an extensive auxiliary network S_{AUX} , de-anonymizing the target network S_{SAN} is difficult

Auxiliary Information

- Auxiliary information is global in nature
  - Many social networking sites overlap one another
  - Facebook, Myspace, Twitter, etc. (correlate)

- Can be used for large-scale re-identification

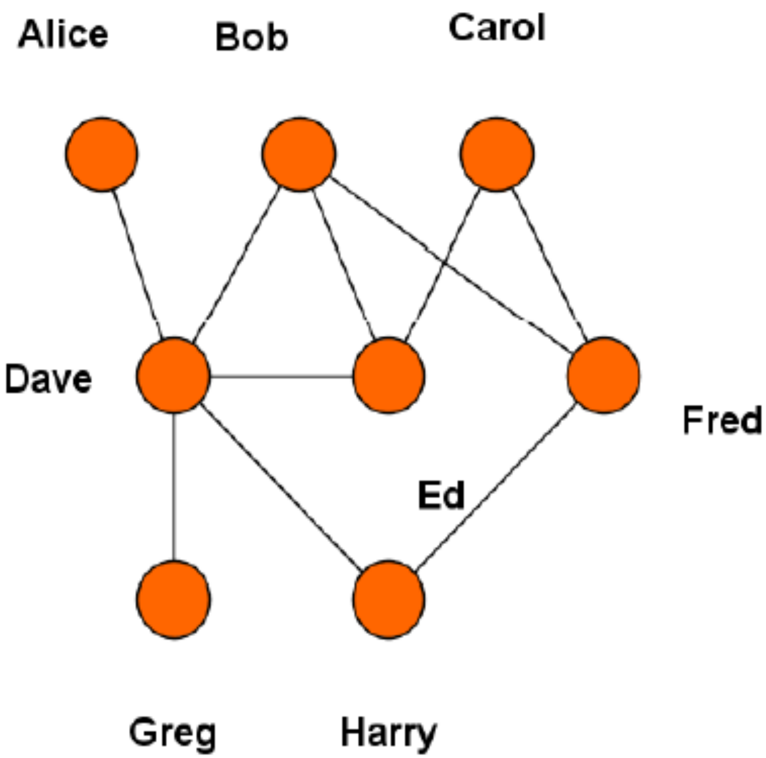
- Feedback based attack
  - Re-identification of some nodes provides the attacker with even more auxiliary information

Individual Auxiliary Information

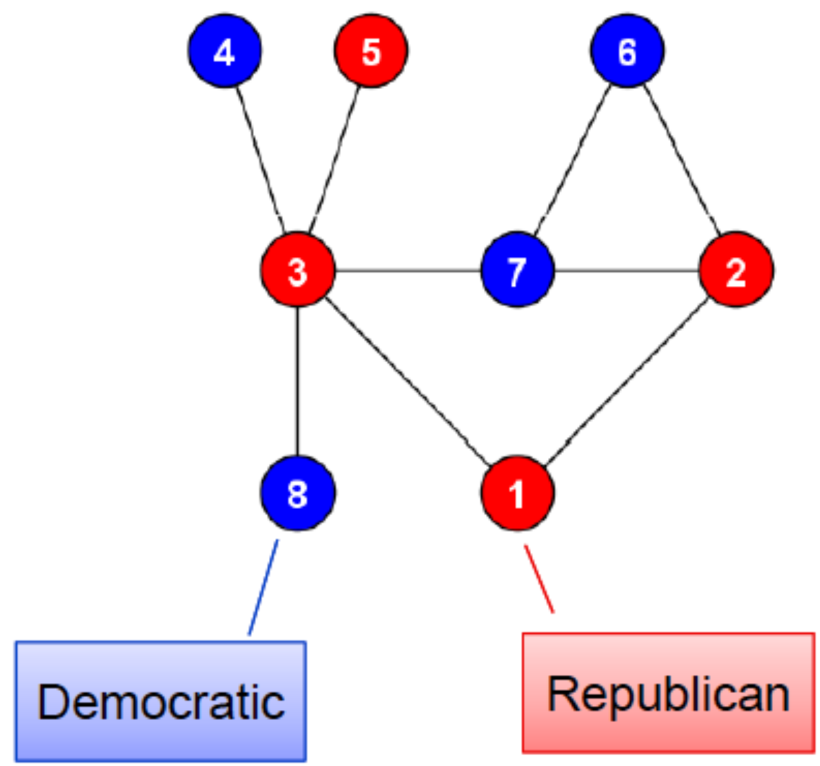
- Assume also that the attacker possesses thorough information about a very small number of nodes on the target network S_{SAN}
- The attacker should be able to identify if those members are also members of his auxiliary network S_{AUX}
- Question at hand: can this information be used in any way to learn sensitive information about *other* members of S_{SAN} ?

Example

Auxiliary information, G_{src}
(a public crawl, e.g., Flickr)



Anonimized graph, G_{tar}
(anonimized export, e.g., Twitter)



De-anonymization

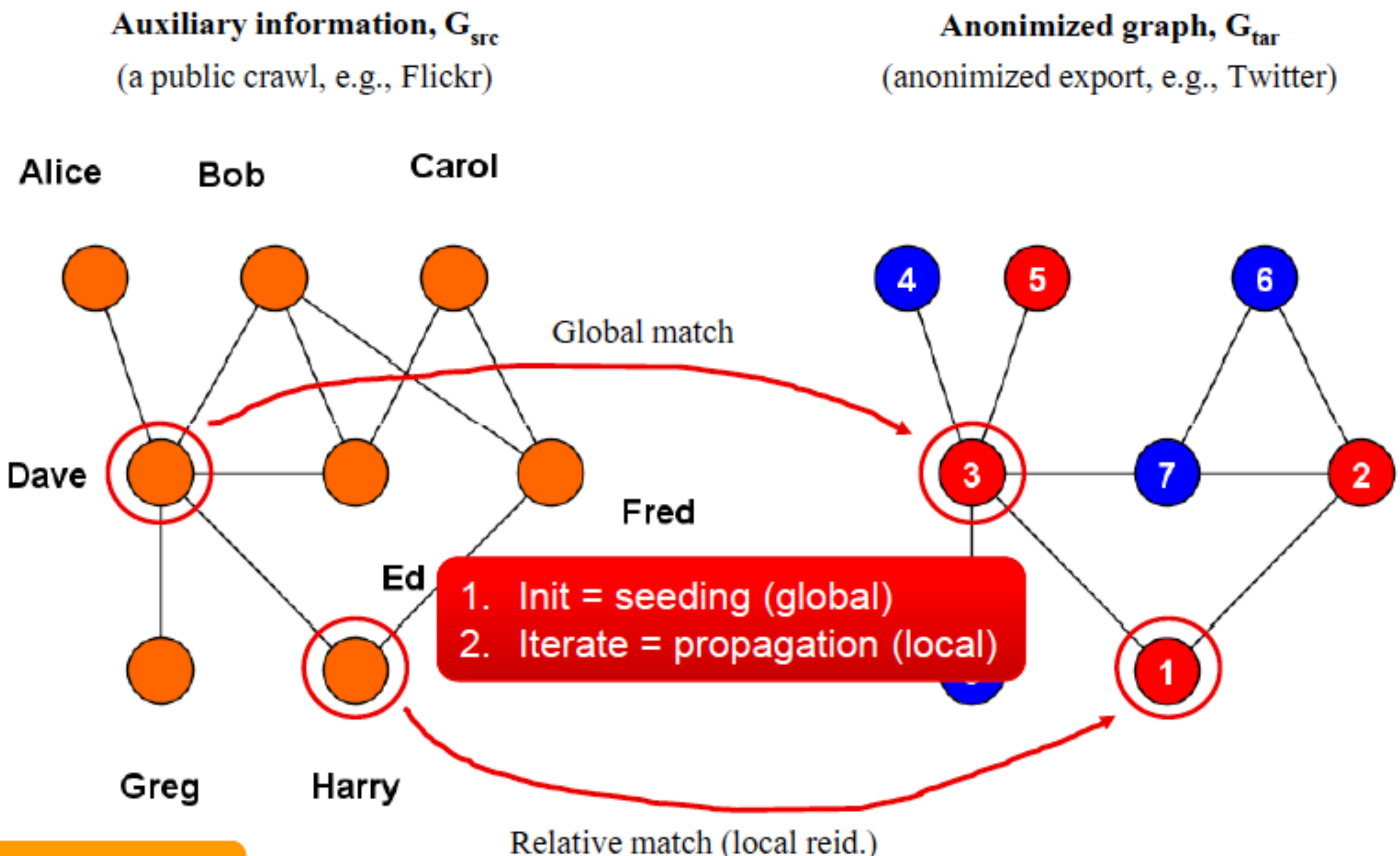
- Two Stages

1. Seed Identification
   – attacker identifies a small group of "seed" nodes which are present in both the anonymous target graph and the attacker's auxiliary graph, and maps them to each other

2. Propogation
   – a self-reinforcing process in which the seed mapping is extended to new nodes using only the topology of the network, and the new mapping is fed back to the algorithm.

- Result is a huge mapping between subgraphs of the auxiliary and target networks which re-identifies (de-anonymizes) those mapped nodes.

De-anonymization



Narayanan & Shmatikov, 2009

Tackling Structural De-anonymization

- Data Sanitization
- Identity Separation

Data Sanitization

- Data sanitization is changing the graph structure in some way to make re-identification attacks harder.
- Most rely on simple removal of identifiers
- Others inject random noise into the graph
- As we said with k-anonymization, trying to make different nodes look the same is not realistic.

Identity Separation