Privacy, Security and Professional Ethics

UBCO Master of Data Science – DATA 553

SOME SLIDES FROM DSCI 541- DR. ED KNOR – MDS VANCOUVER
SOME SLIDES FROM DATA 553 – DR FATEMEH FARD – MDS OKANAGAN

What is Privacy?

General definition : “Right to be let alone” (1890 - S.D Warren and L. Brandeis)

- Freedom of intrusion
- Freedom of surveillance

The state of being free from public attention.

What is Privacy?

Information Privacy - Alan F. Westin

“Privacy is the claim of individuals, groups or institutions to determine for themselves when, how, and to what extent information about them is communicated to others”

- Control of information about oneself
- A right to be forgotten

The right to be forgotten is distinct from the right to privacy. The right to privacy constitutes information that is not publicly known, whereas the right to be forgotten involves removing information that was publicly known at a certain time and not allowing third parties to access the information

What is privacy?

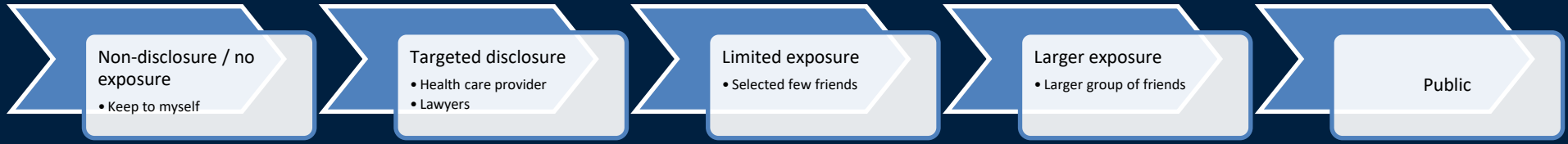
"**Information privacy**, or **data privacy**, is the relationship between the collection and dissemination of data, technology, the public expectation of privacy, legal and political issues surrounding them."

"Privacy concerns exist wherever personally identifiable information or other sensitive information is collected, stored, used, destroyed or deleted. "

Improper or non-existent disclosure control can lead to privacy issues.

https://en.wikipedia.org/wiki/Information_privacy

Degree of Privacy



Privacy policy

Privacy policy

The OIPC has launched a privacy-protective web analytics program which is hosted on our website and will be used to record non-identifiable information about your site visit. The data collected will be stored on our web server and not shared with third parties.

View our Privacy Policy.

Privacy @ Oracle
Oracle General Privacy Policy

Google Privacy & Terms

Overview | Privacy Policy | Terms of Service

Canada Privacy right related Laws

Federal Laws:

- Privacy Act
  - Relates to a person's right to access and correct PI that the Government of Canada collects, uses and holds
  - Only applies to federal government institutions
- PIPEDA (Personal Information Protection and Electronic Document Act)
  - Applies to personal information held by private sector organizations that conduct business generally in provinces and territories other than Alberta, BC and Quebec.
- Enforced by the Office of the Privacy Commissioner of Canada

BC Laws:

- FIPPA
- PIPA
- Enforced by the Office of the Privacy Commissioner of BC

FIPPA

FIPPA: Freedom of Information and Protection of Privacy Act

FIPPA provides a right of access to records held by public bodies and regulates how public bodies manage personal information

Sample Privacy Notification for Direct Collection:

Your personal information is collected under the authority of section 26(c) of the *Freedom of Information and Protection of Privacy Act* (FIPPA). This information will be used for the purpose of evaluating your application for admission to UBC. Questions about the collection of this information may be directed to admissions@ubc.ca.

Sample Privacy Notification for Indirect Collection:

UBC may contact your references to determine your suitability for employment. Your personal information is collected under the authority of section 26(c) of the *Freedom of Information and Protection of Privacy Act* (FIPPA). This information will be used for the purpose of evaluating your application for employment. Questions about the collection of this information may be directed to hr@ubc.ca.

https://universitycounsel.ubc.ca/files/2020/01/Fact-Sheet-Collecting-Personal-Information.pdf

FIPPA

FIPPA: Freedom of Information and Protection of Privacy Act

Purpose: make public bodies more accountable to the public and protect personal privacy

Data residency

A public body must protect personal information in its custody or under its control by making reasonable security arrangements against such risks as unauthorized collection, use, disclosure or disposal.

Public bodies are required to conduct a PIA for all new or substantially modified projects. A “project” refers to any system, process, program or activity that supports its business.

https://privacymatters.ubc.ca/privacy-impact-assessment

Privacy Impact Assessment (PIA)

A privacy impact assessment (PIA) is an analysis of how personally identifiable information (PII) is handled to ensure compliance with appropriate regulations, determine the privacy risks associated with information systems or activities, and evaluate ways to reduce the privacy risks.

- Includes
  - what the project is about
  - Data collected and its sensitivity (risk if disclosure occurs)
  - Impact on current system, integrations (record linkages), transmission (cloud/intranets)
  - Safeguard put in place

Demonstrate the level of security is appropriate to the risks associated

https://www.canada.ca/en/revenue-agency/services/about-canada-revenue-agency-cra/protecting-your-privacy/privacy-impact-assessment/canada-recovery-benefits.html

FIPPA

Privacy Breach Notification

Public bodies that suffer a “privacy breach” (broadly defined as theft or loss, or the unauthorized collection, use or disclosure, of personal information in the custody or control of a public body) must, without unreasonable delay, notify affected individuals and the Commissioner if the privacy breach “could reasonably be expected to result in significant harm to the individual”.

“Significant harm” is broadly defined and includes identity theft, significant humiliation, significant damage to reputation or relationships, significant loss of employment, business or professional opportunities, significant financial loss, and significant negative impact on a credit record.

There are limited exceptions to the obligation to notify affected individuals.

PIPA

PIPA : Personal Information Protection Act

PIPA governs the collection, use and disclosure of personal information by private organizations.

We will inform our owners and residents of why and how we collect, use and disclose their personal information, obtain their consent where required, and only handle their personal information in a manner that a reasonable person would consider appropriate in the circumstances.

This Personal Information Protection Policy, in compliance with PIPA, outlines the principles and practices we will follow in protecting owners' and residents' personal information. Our privacy commitment includes ensuring the accuracy, confidentiality, and security of our owners' and residents' personal information and allowing them to request access to, and correction of, their personal information.

https://www.visoa.bc.ca/wp-content/uploads/2013/06/Privacy-Policy-sample-2014.pdf

PIPA privacy rules

BC rules for Personal Information Protection Act (PIPA) is based on ten principles of privacy protection:

- Be accountable
- Identify the purpose
- Obtain consent
- Limit collection
- Limit use, disclosure and retention
- Be accurate
- Use appropriate safeguards
- Be open
- Give individuals access
- Provide recourse

https://www2.gov.bc.ca/gov/content/employment-business/business/managing-a-business/protect-personal-information/principles

Goals

Preserve privacy

Provide useful data to researchers, governments, consumers, etc. for a purpose.

Health care:

- Heath benefits, reduce costs, new drugs
- Faster to market
- Avoid treatment that are unlikely to work

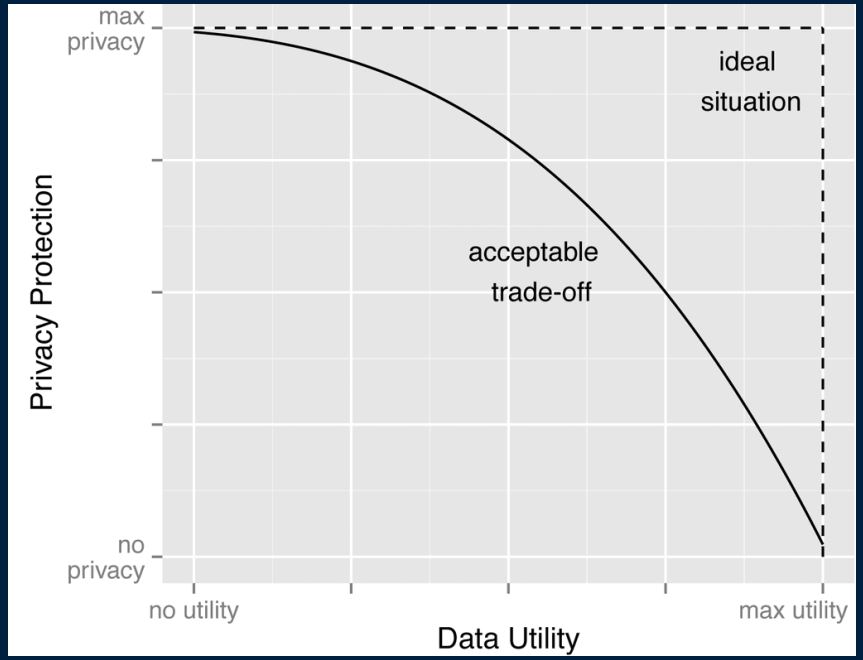


Figure from: Khaled El Emam and Luk Arbuckle. *Anonymizing Health Data: Case Studies and Methods to Get You Started. O'Reilly, 2013.*

Anonymity and Pseudonymity

"There are two different types of anonymity. The first is complete anonymity: a letter without a return address, a message in a bottle, a phone call in a world without Caller ID or phone tracing." [Schneier]

- Anonymity is irreversible. All links between an individual and his/her record are removed. You won't be able to determine the owner.

"The second type of anonymity is more properly called pseudonymity. … It doesn't care who you are, only that you're the same person that deposited the money last week." [Schneier]

- e.g., a pseudonym, userid, or non-personally identifiable e-mail address used for applications like Yelp, Piazza, TurnItIn, GoogleDocs, etc.

De-identification and Re-identification

De-identification: The process of stripping the identifying information from a record, and taking care to make sure that the quasi-identifiers cannot identify the person.

- The original record remains, via a trusted third party.
- Map a record to its owner, possibly by using a one-way hash function h(k):
  - Susan h("Susan") = 4e17ba36d2…
  - Andrew h("Andrew") = 76a83d17ff3…
- Or, just maintain a simple map:
  - Susan => record # 15050
  - Andrew =>record # 5981

Re-identification: Taking a de-identified record and reverse-mapping it to its owner (a link exists via a trusted third party)

- Needs to be authorized by the appropriate privacy board

Basic Form of the Data

A database table (or Excel spreadsheet, etc.) has these disjoint classes of attributes:

1. Explicit (or Direct) Identifiers — **Unique**
   - e.g., name, patient number, social insurance number, address(es), phone number(s), e-mail address(es)
2. Quasi-Identifiers — **Non-Unique**
   - e.g., age, date of birth, gender, occupation (especially unusual occupations), city, postal code (especially in Canada), zip code
     - Be up front with subjects and participants through end-user agreements.
3. Sensitive Attributes—Researchers need these!
   - e.g., disease, disability, health condition, salary
4. Non-Sensitive Attributes
   - all others

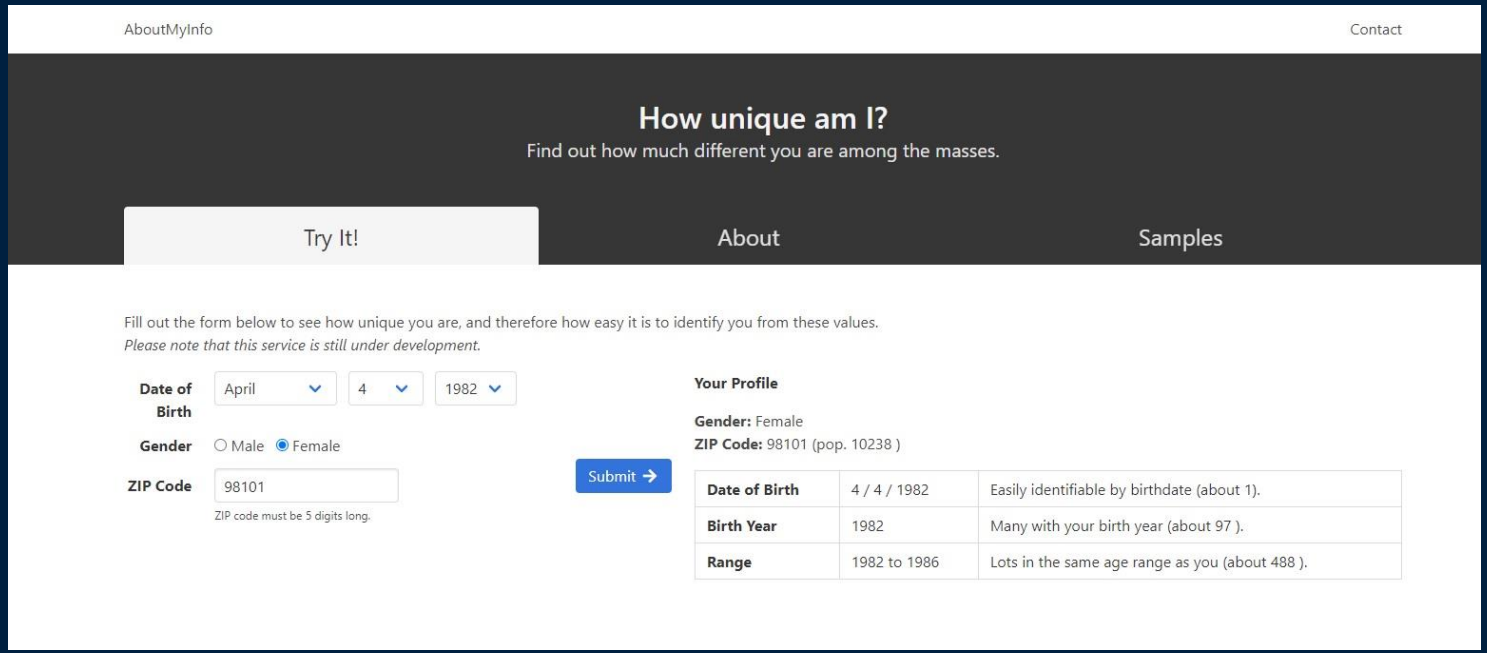
The Problem

Even if we remove the explicit identifiers, the quasi-identifiers when *linked* (in the database world: *joined*) with other information may be sufficient to either identify the individual or narrow the search down to a very small subset of individuals.

- e.g., linking to public voting lists which have zip code, birthdate, and gender
- e.g., if you know where someone (e.g., celebrity, your boss) lives, plus the fact that they went for medical treatment at such a facility at such a time …

For example, a 2002 study by Sweeney showed that 87% of the US population could be inferred from publicly-available information, namely just: 5-digit zip code, gender, and date of birth.

The Problem



https://aboutmyinfo.org

Latanya Sweeney's Finding

In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees

- GIC has to publish the data:

GIC(**zip, dob, sex**, diagnosis, procedure, ...)

This is PRIVATE data?!

Slide source: https://courses.cs.washington.edu/courses/cse544/12sp/lectures/lecture19-privacy.pdf

Latanya Sweeney's Finding cont.

Voter list:

voter registration list for Cambridge Massachusetts:

VOTER(name, party, ..., **zip, dob, sex)**

This is PRIVATE data?!

Slide source: https://courses.cs.washington.edu/courses/cse544/12sp/lectures/lecture19-privacy.pdf

Latanya Sweeney's Finding cont.

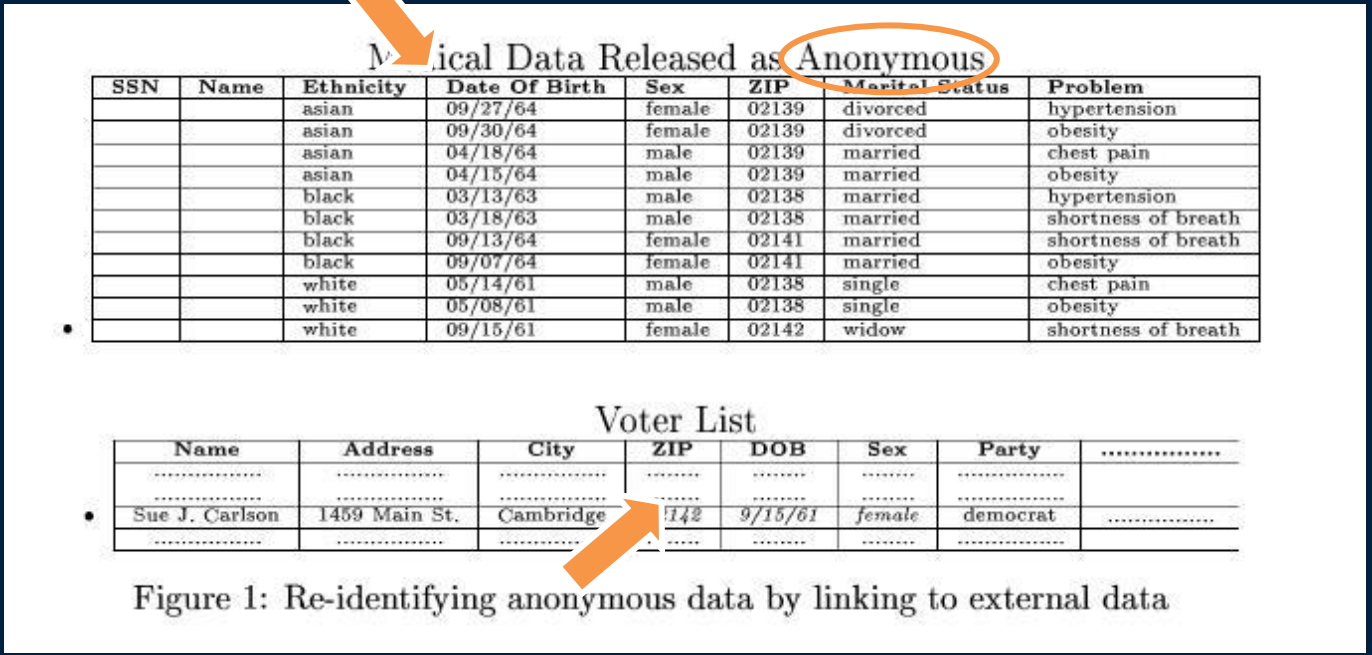
zip, dob, sex

- William Weld (former governor) lives in Cambridge, hence is in VOTER
- 6 people in VOTER share his **dob**
- only 3 of them were man (same **sex**)
- Weld was the only one in that **zip**
- Sweeney learned Weld's medical records !

Slide source: https://courses.cs.washington.edu/courses/cse544/12sp/lectures/lecture19-privacy.pdf

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	142	9/15/61	female	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

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

Public voter dataset

HIPAA Privacy Rule

Regulation designed to protect personal information and data collected and stored in medical records: "Under the safe harbor method, covered entities must remove all of a list of 18 enumerated identifiers and have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information."

“The identifiers that must be removed include direct identifiers, such as name, street address, social security number, as well as other identifiers, such as birth date, admission and discharge dates, and five-digit zip code. The safe harbor requires removal of geographic subdivisions smaller than a State, except for the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same initial three digits contains more than 20,000 people. In addition, age, if less than 90, gender, ethnicity, and other demographic information not listed may remain in the information. The safe harbor is intended to provide covered entities with a simple, definitive method that does not require much judgment by the covered entity to determine if the information is adequately de-identified."

GDPR

The General Data Protection Regulation 2016/679 is a regulation in EU law on data protection and privacy for all individuals within the European Union and the European Economic Area.

It also addresses the export of personal data outside the EU and EEA areas.

- GDPR applies to any entity (any person, business, or organization) that collects or processes personal data from any person in the European Union.

Trying to Solve the Problem

Suppose we leave out the explicit and quasi-identifiers, and substitute an anonymous version of the quasi-identifiers (and/or sensitive attributes), instead:

- Individual records might be aggregated (generalized into a group in which the individual cannot (practically) be identified, vs. individual records)
- Randomization might be added, while preserving the statistical distribution
- Some "noise" might be added
  - The noise added to sensitive value *s* is given by *s* + *r*, where *r* is a random value drawn from a known distribution.
- Some outliers may be removed
  - What is an outlier?

Confounding Issues

Multiple publishers

- Multiple organizations release anonymized data, but maybe someone can link them together and perhaps identify someone

Multiple releases

- Sometimes more data is released than should be, and it's corrected later on. Or, maybe additional tables become publicly available.
- Sometimes new anonymized data is added to a database that may reveal information that changes the statistical distribution. Because of the new data, someone may be able to infer individuals from the "delta" (change).

Repeated querying to probe the data, even if it only reveals the data about groups of 5 people

- Ask enough such queries, and you may be able to infer …

Multiple querying example

(physician, patient, medication)

Query:

listing the patients seen by each physician: (physician, patient)

medications prescribed by each physician: (Physician, medication)

Not safe: patients with their prescribed medications may be sensitive because medications typically correlate with diseases

Solution: restrict such queries

Read more:
https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf

Privacy Threats

A privacy threat occurs when an “attacker” can link a record owner to:

- A record in a published table
- A sensitive attribute (column, field) in a published table
- The table itself
  - e.g., just knowing that someone is in a table of HIV records …

A table is said to be privacy-preserving if it can prevent an attacker from making these linkages.

Record Linkage

If the victim's quasi-identifier matches the quasi-identifier being published, the victim is at risk of being identified.

- Perhaps there are only a small number of possible people that would match the released information.
- If so, perhaps the victim could be identified with some *additional* public information.

Important Note: Data fields that are less identifying, such as date of attendance, are usually not pseudonymized so could be at risk of such inference attack.

Record Linkage (cont.)

How do we protect against freeform text data (e.g., a doctor's comments in a field that are probably fairly specific to an individual)?

- Replace the text with coded data (e.g., SNOMED = standard nomenclature).

What about gene sequence data?

- How do you anonymize that?
  - Fraser & Willison: "Even a few dozen gene markers may provide enough data to uniquely identify an individual from a genetic sample. The forensic use of such gene sequence data makes it at least as privacy invasive as a complete set of fingerprints."
- A "solution": Limit its use and distribution, encrypt it, and use access control and physical security

What about diagnostic images (e.g., coded DICOM data)?

Solution: Pseudonymization and Anonymization

The legal distinction between anonymized and pseudonymized data is its categorisation as personal data.

Pseudonymous data still allows for some form of re-identification (even indirect and remote), while anonymous data cannot be re-identified.

Anonymization: the data is scrubbed for any information that may serve as an identifier of a data subject.

Pseudonymization: does not remove all identifying information from the data but merely reduces the linkability of a dataset with the original identity of an individual (e.g., via an encryption scheme).

Source: https://gdpr.report/news/2017/11/07/data-masking-anonymisation-pseudonymisation/

k-Anonymity

k-anonymity means that the quasi-identifier (*qid*) must appear in at least *k* records in the table.

- e.g., There may be *k* people in the table having the same birthdate and gender. Thus, you're not sure if a friend of yours, for example, is actually the person in the table.

This means:

- The probability of linking an individual to a specific record in the table is at most 1/*k*. Mathematically, these *k* records form an equivalence class.

We call it a *k*-anonymous table.

k-Anonymity (cont.)

The more attributes that make up a quasi-identifier, the harder it is to distinguish individuals.

- However, this may also make it easier for an attacker to strength his/her case that a given individual is actually *in* the table.

You may have to distort more of the data (e.g., go higher up the taxonomy) if you use more attributes; but, this may make the data *less useful*—a trade-off.

If a given individual has multiple records in the same table, then the group of *k* records offers less protection to that individual.

k-Anonymity (cont.)

But, what about the sensitive attributes? How do they relate to *k*-anonymity?

If many of the *k* records share a similar sensitive attribute, this may compromise privacy.

- e.g., Suppose *k* = 5, and suppose 4 of 5 individuals have the same disease. This means there is an 80% chance that a given individual in this *qid* group also has that disease.
- Obviously, it's even worse if all 5 individuals have the same disease!

Disclosure of Record-Level Data

Some forms of Data Reduction

- Removal of explicit identifiers
- *k*-anonymity
- Generalization (or aggregation) of data
  - May go hand-in-hand with k-anonymity
  - Date of birth can be rounded to year of birth
  - Use durations or intervals (e.g., 15-day waiting time, 1st quarter of 2017)
  - Postal code may be a giveaway (e.g., a single apartment building may have a unique postal code) … can reduce this to the first 3 alphanumeric characters
  - What do we do with rare medical conditions?
- Sampling

Disclosure of Records (cont.)

Some forms of Data Modification

- Adding “noise” to the data (e.g., +/- integers up to 5, list birthday within a 6-month range)
- Replace names and ID numbers with random strings
- Data swapping (two similar records could have some data swapped)

Some forms of Data Suppression

- Removal certain cells (values) in a row
  - Useful when encountering outliers
- Remove of certain rows altogether
  - Useful when there’s a non-trivial probability of identification

Some forms of Pseudonymization

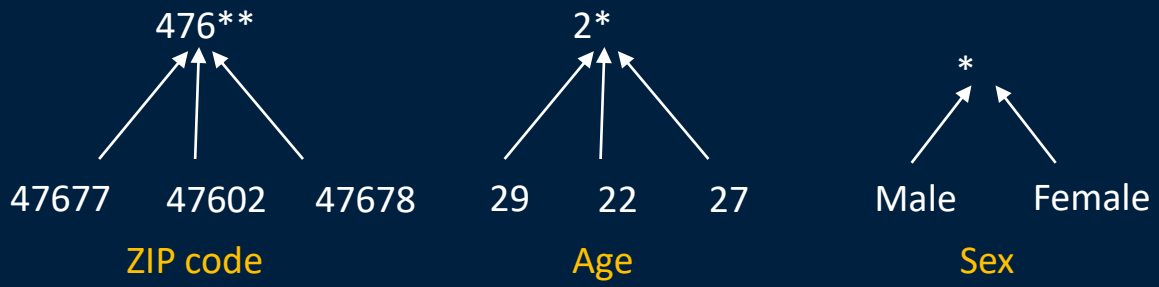
- Encryption
- Hashing (e.g., via look-up table to maintain the link, privately)

Generalization

Generalization is a form of data reduction.

Consider a taxonomy that goes from very specific items (e.g., “541 ml. Coke”) at its lowest (finest granularity) level, up the tree to “Coke”, then to “Coke products”, then to “cola”, then to “soft drinks”, and then to “beverages”—you pick the granularity

Generalization: Replace quasi-identifiers with less specific, but semantically consistent values.



Achieving k-Anonymity

Generalization

- Replace individual quasi-identifiers with less specific values until we get at least *k* identical values
- We can partition ordered-value domains (e.g., numeric values) into intervals (like histograms)

Note that we can perform multidimensional generalization

- e.g., <engineer, male> and <engineer, female> could generalize to <engineer, any_sex>, <professional, female>, etc.
- This allows us to generalize only the *qid* groups that violate a specified threshold.
- But, things could get tricky in the presence of additional information (e.g., linkages)—a data exploration problem.

Achieving k-Anonymity (cont.)

Cell Suppression

- A form of data reduction
- Generalization might cause too much information loss.
  - e.g., due to the presence of outliers
- We replace cells with a special value (or perhaps leave them blank).
- In some cases, we suppress the whole record.
- Suppression is the opposite of *disclosure*.

Lots of algorithms about this in the database and statistics literature

- We want to produce *useful* anonymizations.
- Note that the publisher of the data may not know the purpose of the studies.

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*}

Example of Generalization

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

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

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

http://www.cs.cornell.edu/~shmat/courses/cs5436/anonymization.pdf

Example of Generalization (cont.)

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

!!

The released table is 3-anonymous.
If the attacker knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record; but, what about Bob?

http://www.cs.cornell.edu/~shmat/courses/cs5436/anonymization.pdf

l-Diversity and Attribute Linkage

With *l*-diversity, every *qid* group must have at least *l* "well-represented" sensitive values.

- e.g., Suppose we have one sensitive attribute, and *k*-anonymity. We need to have at least *l* distinct values for that sensitive attribute—within a *qid* group.
- If there are *l* distinct sensitive values and each distinct value occurs once in the group, then the probability of linking a record to a sensitive value is 1/*l*.

Sensitive Attribute Disclosure

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 Doesn't Eliminate Inference Attacks

This does not prevent probabilistic inference attacks because some sensitive values (e.g., HIV) occur much less commonly than others (e.g., cold, flu).

Suppose the victim is in a group that has 3 different sensitive values: cancer, flu, and HIV.

- If the attacker knows that the victim shows no sign of having the flu …
- If persons A and B have a contagious disease, and person C lives with them, this may infer that C also has that disease.

Enforcing a rule that $k = l$ may lead to high information loss.

- e.g., If only 5 people in a table of 1000 records have HIV, and if a sensitive attribute has a binary (yes/no) indicator for HIV, and if at least 1 HIV patient is needed in each *qid* group, then at most 5 groups can be formed.

De-identification and Re-identification

De-identification: The process of stripping the identifying information from a record, and taking care to make sure that the quasi-identifiers cannot identify the person.

- The original record remains, via a trusted third party.
- Map a record to its owner, possibly by using a one-way hash function h(k):
  - Susan h("Susan") = 4e17ba36d2…
  - Andrew h("Andrew") = 76a83d17ff3…
- Or, just maintain a simple map:
  - Susan => record # 15050
  - Andrew =>record # 5981

Re-identification: Taking a de-identified record and reverse-mapping it to its owner (a link exists via a trusted third party)

- May be done for very good reasons … like what?
- Needs to be authorized by the appropriate privacy board

Summary

The public would benefit from the publishing of “sensitive” data.

- But, it has to be done “right”!
  - Privacy-preserving data publishing

Data may have:

1. Explicit Identifiers
2. Quasi Identifiers
3. Sensitive Attributes
4. Non-Sensitive Attributes

Quasi identifiers can be used to link (or identify) individuals, even in the absence of explicit identifiers.

Summary (cont.)

In *k*-anonymity, released data cannot be distinguished from at least $k - 1$ other individuals in the table.

- e.g., The probability of linking an individual to a specific record in the table is at most $1/k$.

Generalization can be performed by providing coarse-grained attributes instead of fine-grained attributes.

- e.g., using a taxonomy of health-care conditions to reach ≥ *k* records having the same "generalized" condition
- This allows us to create a pool of at least *k* similar individuals.

Use cell suppression if there would be too much information loss from generalization

Watch for outliers (anomalies that might give away too much information).

Summary (cont.)

Another approach: Add randomization to data records, yet still be able to produce similar distribution and summary statistics.

With *l*-diversity, every *qid* group must have at least *l* “well-represented” sensitive values.

But this doesn’t eliminate the possibility of probabilistic inference attacks.


THE UNIVERSITY OF BRITISH COLUMBIA