SFU LIBRARY

DATA DE-IDENTIFICATION

SFU Library Data Services
data-services@sfu.ca
http://www.lib.sfu.ca/data

DATA DE-IDENTIFICATION: GOALS

1. Identify personal information, direct and indirect identifiers in data
2. Understand risks of disclosure
3. Methods of de-identification and anonymizing data
4. Strategies and tips for risk assessment and mitigation

DOES YOUR DATA NEED DE-IDENTIFICATION?

Will the data be shared?

- Public or limited access

Is there personally identifiable information (PII)?

- How much?
- Level of sensitivity

What is the risk of disclosure (unauthorized access or release)?

- Potential harm to individuals
- Likelihood of disclosure
- Legal requirements

WHAT IS PERSONAL INFORMATION?

Information about an *identifiable* individual

Direct identifiers: explicitly point to an individual	Inferential/quasi-identifiers: could be combined to identify an individual
• Name • Address • SIN • PHN • Phone number	• Age • Postal Code • Race or Ethnicity • Income • Medical diagnoses

RISKS OF DISCLOSURE

Potential harm to participants

- Sensitive information (e.g. health information, sexual orientation, criminal record) can cause greater harm if disclosed

Legal consequences

- Provincial and federal privacy laws may apply
- Fines, investigations, lawsuits

WHAT INCREASES RISK?

Potential for re-identification of individuals

Links to other public data

- Repeated identifiers, ability to search text present in another dataset for more information
- Mining social networks can connect multiple social media profiles of single user
- Data built on a previous study using same identifiers

Small populations, unique groups

- Too few individuals to ensure anonymity
- Examples: GPS data from an individual's phone; Health statistics for a small city or state

WHAT INCREASES RISK?

Exact values: less specificity decreases risk

- Specific ages instead of age ranges (e.g. 18-24)
- 3-character postal code vs full

Individual profile: combined variables of a single participant

- Example: using the same identifier for an individual across multiple datasets

RISK MITIGATION STRATEGIES

Access restrictions

- Greater access means greater risk
- Share data after certain time period
- Restrict to institution or permission-based
- Balance with requirements from funder or institution, value to other researchers

De-Identification & Anonymizing

- Methods to protect individuals' information in the data

ANONYMIZING STRATEGIES - MASKING

Used for direct identifiers

Redact/remove

Create pseudonyms

- Avoid using the same identifier for an individual across datasets
- Use a secure method: hashing, encryption

Randomized variables

- In numeric data, include random noise while keeping the distribution the same as original

Substitution or Shuffling

- Replace actual values with similar realistic ones
- Swap values between individuals within a database

DE-IDENTIFICATION STRATEGIES

Use for direct or indirect identifiers to prevent de-anonymizing

Generalize values

Supress unique values or one of multiple quasi-identifier fields

Subsample: only release part of a dataset

Aggregate: make summary data available for dataset or only certain fields

HOW MUCH DE-IDENTIFICATION IS ENOUGH?

k-anonymity – common standard for determining acceptable level of anonymity

An individual has *k*-1 other individuals with the same identifying attributes in the data

EXAMPLE

PHN	Age	Postal Code	Language	Gender
3257962234	28	V6J 3A7	English	Male
5864385686	34	V1K 1T6	English	Female
1276590438	27	V6J 0V2	Mandarin	Male
2047893456	24	V5M 0T5	French	Male
1234567892	35	V1K 0B2	Mandarin	Female
5436874328	23	V5M 1B8	French	Female
4321657765	29	V6J 3B2	English	Male
3454544227	29	V5M 3M4	Cree	Female
9678463721	37	V1K 1P2	Arabic	Female
8674563214	27	V5M 4R5	English	Female

EXAMPLE: DE-IDENTIFIED

PHN	Age	Postal Code	Language	Gender
NULL	[20, 30)	V6J	NULL	Male
NULL	[30, 40)	V1K	NULL	Female
NULL	[20, 30)	V6J	NULL	Male
NULL	[20, 30)	V5M	NULL	Male
NULL	[30, 40)	V1K	NULL	Female
NULL	[20, 30)	V5M	NULL	Female
NULL	[20, 30)	V6J	NULL	Male
NULL	[20, 30)	V5M	NULL	Female
NULL	[30, 40)	V1K	NULL	Female
NULL	[20, 30)	V5M	NULL	Female

PHN may also be scrambled in case similar number format is needed in analysis

Summary data optional for masked fields:

- E.g. 40% of participants listed English as their primary language; mean age of 29

EXAMPLE: DE-IDENTIFIED

PHN	Age	Postal Code	Language	Gender
NULL	[20, 30)	V6J	NULL	Male
NULL	[30, 40)	V1K	NULL	Female
NULL	[20, 30)	V6J	NULL	Male
NULL	[20, 30)	V5M	NULL	Male
NULL	[30, 40)	V1K	NULL	Female
NULL	[20, 30)	V5M	NULL	Female
NULL	[20, 30)	V6J	NULL	Male
NULL	[20, 30)	V5M	NULL	Female
NULL	[30, 40)	V1K	NULL	Female
NULL	[20, 30)	V5M	NULL	Female

k-anonymity for $k=3$ on age, gender, and postal code; for any row, there are at least 2 (or $k-1$) others with the same attributes.

Any additional non-identifier fields can now be released with these identifiers.

ASSESSING RISKS

1. Determine whether PII will be collected and whether your data will be shared.
   - Ensure informed consent is obtained from participants – including awareness of disclosure risks
   - Collect only as much PII as is necessary
   - Is sharing your documentation or summarized data sufficient?
2. Identify direct and indirect identifiers in your data
   - How likely is re-identification if no indirect identifiers were removed?
   - How easy is it to de-anonymize the data to identify individuals?
   - Does your data fall into any of the higher risk categories?

ASSESSING RISKS

3. What level of risk is acceptable?
   - Evaluate risk in a "worst case scenario" disclosure – if all data were disclosed, what is the potential level harm?
   - Consider sensitivity of data, number of individuals potentially affected, and extent of sharing
   - Consider potential research and analytical value of data
4. De-identify according to risk assessment.
   - Balance utility and security
   - Analytically useful identifiers should be kept if possible
   - Masking and de-identification methods to remove quasi- and direct identifiers

ANONYMIZATION TOOLS

Amnesia

- https://amnesia.openaire.eu/

sdcMicro R package

- https://cran.r-project.org/package=sdcMicro

ARX

- https://arx.deidentifier.org/

FOR MORE INFORMATION

Email us at data-services@sfu.ca