Data De-identification and Anonymization: An Introduction

Creating safe(r) shareable data

Key concepts

Anonymization and de-identification, identifiers, quasi-identifiers, risk

Anonymization and deidentification

- Anonymisation is the permanent removal of identifying information, with no retention of the identifying information separately.
- Deidentification is the creation of a version of the dataset with the identifying information removed, while identifying information is retained separately
- Permanent retention of identifying information is *not* required under any anticipated changes to Tri-Agency policies.

Identifiers, quasi and otherwise

- Direct identifiers
  - Names, addresses, telephone numbers, social media tags, unaltered photographs of individuals, or any identifiers used by the researchers to link data to one of the above
- Quasi-identifiers
  - Demographic variables that have the potential to be linked with other data sources to violate the confidentiality of participants.
  - E.G. age, gender identity, income, occupation, industry / place of work, geography, ethnic and immigration variables

Quasi-identifiers

- A variable should only be considered a quasi-identifier if an attacker could plausibly match that variable to information from another source
  - Some variables may be used to derive other quasi-identifiers, e.g. community size could be combined with a broader geographic grouping to infer location more precisely
- A set of records that has the same values on all quasi-identifiers is called an *equivalence class*

Risk

- Risk is created when:
  - Variables can isolate individuals in the dataset

  - Identifying information can be matched to persistent information that an attacker may reasonably have access to

    AND
  - Linking the two (research data and outside data) will give the attacker more information than they would otherwise have

Other forms of identifying information

- With internet and social media data, an additional concern is the possibility of identifying an individual by identifying their computer, network location, or data unique to a social media account
  - IP address ranges, logged traffic patterns which could be linked to IP addresses, etc.
  - List of favourite books / movies on a social media account? Yes.
    - See: Guess Who Rated This Movie: Identifying Users Through Subspace Clustering
- Basic issues remain the same.

Non-identifying information

- Survey responses that are not likely to be recognizable as coming from specific individuals or to show up in other databases
- Usually, most questionnaire responses: opinions, ratings, anything measured with Likert scales…
- Temporary measures: resting heart rate after meditation, number of times ate breakfast last week
- Free text responses / comments / transcribed qualitative interviews need to be considered case by case
  - “The library needs more electrical outlets”: not identifying
  - “And when I spoke to my colleagues at the plant about organizing a union …” possibly identifying

Assessing risk and dealing with risk: mathematics and rules of thumb

A few heuristics and an introduction to K-Anonymity

Assessing quasi-identifiers

- Quasi-identifying variables containing groups with small cell sizes (e.g. a religion variable with 6 individual responses of "Buddhism") pose high risk.
- Extreme values (more than 10 children; very high income) pose high risk
- Size of identifiable groups *in the general population* also need to be considered
  - There may be only one person from Winnipeg in your random digit cell phone user survey, but if your survey doesn't narrow it down any further than that, that person is pretty safe
- Contextual information that accompanies the data should also be part of the analysis
  - If it is clear from the context of your research that all your interview subjects worked at a particular tool and die plant in Oshawa, that narrows things down quite a bit

Heuristics – population-based

- Geographic information should not be released if there are less than 20,000 people living in the geographic area (this comes from US Health Insurance Portability and Accountability Act – HIPAA, but similar heuristics are used by Statistics Canada and the Census Bureau)

- If the quasi-identifier values represent less than 0.5% of the population, then these values represent a rare group and response values on the quasi-identifiers should be generalized or the observations in the dataset suppressed. This heuristic is used by Statistics Canada to top code age at 89+, for example.

Rules of thumb – dataset-internal

- Note that category sizes that are small enough to violate confidentiality are also often of limited or no research value

Common-sense / descriptive

- Look at the demographic variables in the dataset and consider describing an individual to a friend using only the values of those variables. Is there any likelihood that the person would be recognizable?
- “I’m thinking of a person living in Toronto who is female, married, has a University degree, is between the ages of 40 and 55 and has an income of between 60 and 75 thousand dollars.”
  - Even if there is only one such person in the dataset, this is not enough information to create risk…
  - **UNLESS** contextual information about the dataset narrows things down further
    - Let’s say you know this is a survey of referees for the OHA…

K-Anonymity

- K-anonymity is a mathematical approach to demonstrating that a dataset is anonymized
- Concept: it should not be possible to isolate fewer than K individual cases in your dataset based on any combination of identifying variables, where K is an integer set by the researcher (generally not less than 5 and may be higher)
  - That is, a record cannot be distinguished from K-1 other records.
- Several software programs in development to automate checking this; many I have tested have not worked very well on real datasets
- Can be done using cross-tabulations (SPSS, Stata etc.) for limited number of variables

K-Anonymity

- For very large numbers of variables with large numbers of categories this may be a challenge
- But why are you collecting all that demographic information? What do you think this is, the census?
  - Don’t collect potentially identifying information that is not actually needed
- K-anonymity may be overkill depending on dataset-external factors, as well as being time-consuming to assess
  - There are various open / academic software packages that attempt to automate checking datasets for K-Anonymity. Amnesia is one such tool and will be discussed in the second half of this webinar.

Dealing with quasi-identifiers identified as risky

- Usual strategies:
  - Top-coding of outliers (e.g. age of 80+)
  - Grouping into categories e.g. age in 10 year increments
  - Deletion of outliers that cannot be grouped into any existing categories

Other strategies

- Perturbation / random noise
  - Adding or subtracting a random number following a Gaussian distribution from a numeric variable e.g. income.
  - Maintains overall distribution of data but changes the error term in the data in a way that is opaque to the researcher
- Value substitution
  - Taking a value from one record and exchanging it with that of another record
  - Maintains the **univariate** distribution of values in the dataset; need to trust the algorithm that no other errors are introduced. May change bivariate / multivariate distributions.
- Some statisticians love these things, but they have not gained traction among researchers

Case studies

Some simple examples

Case: historical government survey

- This survey data file had not been de-identified by the survey firm employed by the government agency (which was the usual practice)
- Primary issue was presence of a number of geography variables
- Contextual information: the survey was a general survey of the Canadian population

Procedure

- Removed census geography, area codes, grouped provinces into regions where necessary
- Dataset contained indirect community identifiers such as community size (fairly precise); dropped these as well
- Regrouped the categories on demographic variables (education, employment, etc.)
- We were able to use similar published surveys from the same agency that had undergone correct review as a model for how to do the grouping
- Verified by cross-tabulating demographics using a statistical package and reviewed separately by a second librarian

Case: extract from a cardiology program database.

- Variables under consideration: Marital Status, Ethnicity, Religion, Education
- Contextual information: known region, known age range

Marital status

Divorced	188
Married	1,248
Separated	11
Single	149
Widowed	110

Suggested grouping:
Living with spouse Yes / No.
(Practically speaking that is what mattered for the analysis.)

Ethnicity – 7 categories, most small.

Final grouping: retained African-American as that was a subgroup of interest. Researchers wanted to use “Hispanic” grouping but only 20 cases.

African American	114
Caucasian	1,434
Other	165

Education: original had a large set of categories, some quite infrequent / strange

High School or Less	214
Some Postsecondary	520
Postsecondary Degree	870

I once encountered a dataset with an education category for "Attended Swedish normal school."
It was not a Swedish dataset.

Notes

- Cardiology dataset:
  - If the researchers follow my recommendations, the cardiology dataset would end up with no equivalence classes among the quasi-identifier variables smaller than 20 (k = 20)
  - It was possible to do this by hand because of the relatively small number of demographic variables present
  - Given that this dataset is of a small known population the data will not be widely shared – this was just about getting it to the point where a broader research team could work on it.
- Historical government survey
  - This survey dataset was more complex but also much larger and was a general population survey, so was of lower risk to begin with
  - We also had a model to follow that had been used on many similar datasets done by the same organization

Final observations

- Manual de-identification of data is difficult and time consuming, and the difficulty increases exponentially with the number of potentially identifying variables present.
- I downloaded and tested a number of (free) software packages that were developed to assist with de-identification. I found that for someone at my level of knowledge, all were too complex, buggy or poorly documented to be of much help, though the package Amnesia (which you will be hearing about shortly) seemed promising.
- Software aimed at the general academic survey researcher should not assume special knowledge in the field of data de-identification. Software aimed at professional survey firms can probably assume a much higher level of training.