12-2-2022

Safe Sharing for Sensitive Data

Kristi Thompson
Western University

Safe Sharing for Sensitive data

DREAM OR MYTH?

Western
UNIVERSITY · CANADA

Land Acknowledgement

I am joining you today from Western University in London, Ontario, which is occupying the traditional lands of the Anishinaabek, Haudenosaunee, Lūnaapéewak, and Attawandaron peoples

What is Research Data Management?

The organization and maintenance of data throughout the research process

Incudes

- Setting up plans and processes before starting data collection
- Keeping track of, documenting, and backing up data during the research project
- Destroying, archiving or publishing data after the project has completed

If you collect or use data you are doing RDM

- Possibly not standards-compliant, Tri-Agency approved RDM

Tri-Agency Data Management Policy

Policy includes 3 requirements:

- Institutions: Institutional Strategy.
  The Strategy will explain how researchers will be supported in coping with the next two requirements
- Researchers: Data Management Plans.
  Data management plans will be required for selected funding opportunities.
- Researchers: Data Deposit.
  Researchers will be required to deposit data and code directly associated with research publications into an approved repository. Note that this is not the same as requiring open sharing.

Requirements are being rolled out at the level of the individual grant.

Tri-Agency on Data Sharing

This is what the Tri-Agency policy says about data sharing:

- Grant recipients are not required to share their data. However, the agencies expect researchers to provide appropriate access to the data where ethical, cultural, legal and commercial requirements allow…

"Provide appropriate access" does not mean sharing all data openly, and the policy has many exceptions baked in. But most current academic and journal repositories are designed to make open sharing the most convenient option, and the policy definitely encourages sharing with the implication that sharing may help you get funds.

I'm approaching this as a research data management librarian – someone expected to help researchers comply with this policy.

Are researchers expected to understand how and when to safely share data to comply with these expectations? Are data curators? Research ethics boards?

Background and key concepts

IDENTIFIERS, QUASI-IDENTIFIERS, RISK

Direct Identifiers

Any information collected by the researcher that places study participants at immediate risk of being reidentified

Full or parts of: Names, addresses, telephone numbers, or any identifiers used by the researchers to link data to one of the above

Detailed geography (areas containing less than 20,000 people is a rule of thumb - HIPAA)

IP addresses and other information that may be associated with a computer

Exact dates linked to individuals or events are highly identifying

HIPAA recognizes 18 personal identifiers that will qualify data as personal health information; the BMJ compiled a list of 28 based on multiple international research guidelines

Indirect or Quasi-identifiers

Quasi-identifiers are commonly thought of as demographic variables and socio-economic variables that have the potential to be linked with other data sources to violate the confidentiality of participants, or to be recognized by a person acquainted with the survey respondent.

- Specific examples include age, gender identity, income, occupation, industry / place of work, geography, ethnic and immigration variables

Potentially, membership in specific organizations, use of specific services

Variables that relate to geography in any way need to be treated with extreme caution

- Potential community identifiers can include features like presence of a university hospital or international airport
- E.G. variable giving distance to nearest emergency department
- Need to be considered alongside any contextual information about the dataset

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

Risk – a technical definition

Risk is created when:

- Variables can isolate individuals in the dataset
- Identifying information can be matched to persistent information somewhere else that an attacker may reasonably have access to

A set of records that has the same values on all quasi-identifiers is called an *equivalence class*

An equivalence class of one corresponds to an individual who is unique in the dataset on some combination of characteristics. Such a person may be at risk of being identified.

- This person is called a *sample unique*. If your survey is a complete sample of some population, this person is also a *population unique*.

Assessing and dealing with risk: statistical disclosure risk assessment

AN INTRODUCTION TO

K-ANONYMITY

Assessing quasi-identifiers

- Quasi-identifying variables containing groups with small numbers of respondents (e.g. a religion variable with 3 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

Common sense (can only take you so far)

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…

BUT – consider unusual combinations of variables – let's say someone belonged to the 18 and under age group and also responded that they are widowed.

How do you figure this out without needing to know every single combination in the data?

Keep in mind that due to the miracle of multiplication, a simple dataset – say, 5 demographics with 2-5 categories each – can easily generate several hundred possible combinations (e.g. 3*2*3*5*4 = 360)

k-anonymity

k-anonymity is a mathematical approach to demonstrating that a dataset is anonymized

- First proposed by computer scientists in 1998 and has formed the basis of formal data anonymization efforts since then

Concept: it should not be possible to isolate fewer than k individual cases in your dataset based on any combination of matching characteristics

That is, a record cannot be distinguished from K-1 other records in its equivalence class.

k is just a number set by the researcher; three and five are commonly used

Equivalence classes and “data twins”

It should not be possible to isolate fewer than k individual cases in your dataset based on any combination of identifying variables

Cases 1, 6 and 13 form an equivalence class with k=3

- Each case in the equivalence class has 2 “data twins”

Case 14 has no data twins – it is a sample unique

A dataset’s k is the size of the smallest equivalence class in the dataset – in this case 1.

ID	Gender	AgeGrp	EthnicGrp
1	M	25-30	1
2	F	16-24	1
3	M	25-30	2
4	M	16-24	1
5	F	31-45	1
6	M	25-30	1
7	F	16-24	1
8	F	31-45	1
9	F	31-45	2
10	M	25-30	2
11	M	16-24	1
12	F	25-30	1
13	M	25-30	1
14	F	16-24	2
15	F	31-45	1

Data reduction – global reduction and local suppression

Global data reduction

- Grouping into categories e.g. age in 10 year increments
- For already categorical variables, merging into larger groups
- Complete removal of risky variables from the dataset

Local suppression

- Deleting individual cases or responses
- For example, a member of the ’17 and under’ age group who responded ‘widowed’ might have their response to the marriage question deleted as an alternative to further recoding the otherwise non-risky variables of AgeGroup and MaritalStatus

By looking at frequencies and creating bivariate tables of variables, it is possible to single out the riskiest categories on variables and regroup / suppress them as a prelude to checking k-anonymity, and then look at equivalence classes to find remaining risky cases and fix them

Checking k-anonymity

Stata statistical language:

```
egen equivalence_group= group(var1 var2 var3 var4 var5)
* create a variable to count cases in each equivalence group
sort equivalence_group
by equivalence_group: gen equivalence_size =_N
tab equivalence_group if equivalence_size < 5, sort
```

R statistical language

```
library('plyr’)
# Figure out what equivalence classes there are, and how many cases in each
equivalence class.
dfunique <- ddply(df, .(var1, var2, var3, var4, var5), nrow)
dfunique <- dfunique[order(dfunique$V1),]
View(dfunique)
```

The UK Anonymisation Network Anonymization Decision-Making Framework, appendix B has code for doing this in SPSS.

~~Guaranteed~~ data anonymization

k-anonymity is intended to be a form of guaranteed data anonymization and is often described as such.

It guarantees that every person in the anonymized data will be indistinguishable from k-1 data twins.

However…

Research participants are not generally told that no one will know which line of the data file holds their confidential information. They are told their answers to research questions will be kept confidential.

Attribute disclosure

INTRODUCING L-DIVERSITY AND FRIENDS

Attribute Disclosure

Cases 1, 6 and 13 still form an equivalence class with k=3. So even if you know which people in this survey population match those characteristics, you can't tell which person matches which case

BUT

They all answered a particular question (about whether their workplace should unionize) the same way

You now know how all three of them answered this question. Confidentiality had been violated.

ID	Gender	AgeGrp	EthnicGrp	Unionize
1	M	25-30	1	**Y**
2	F	16-24	1	N
3	M	25-30	2	N
4	M	16-24	1	Y
5	F	31-45	1	Y
6	M	25-30	1	**Y**
7	F	16-24	1	N
8	F	31-45	1	Y
9	F	31-45	2	Y
10	M	25-30	2	N
11	M	16-24	1	Y
12	F	25-30	1	Y
13	M	25-30	1	**Y**
14	F	16-24	2	N
15	F	31-45	1	Y

ℓ-diversity and friends

Extensions of *k*-anonymity, including *p*-sensitive k-anonymity and ℓ-diversity, have been proposed to deal with attribute disclosure; they all involve rules around what values the attributes within an equivalence class should have

Example: one of the simpler variants, called distinct ℓ-diversity

- A dataset satisfies distinct ℓ-diversity if, for each group of records in an equivalence class (matching on all their quasi-identifiers) there are at least ℓ different responses for each confidential variable
- So for our workplace survey, every group of data twins would have to contain both yes and no answers to the "unionize" question, since two would be the maximum possible value for ℓ for this question
- And this would have to be true for some value of ℓ for every confidential answer in the dataset

Imagine a typical survey dataset with dozens of questions, each of which needs to be considered for ℓ -diversity for each equivalence class…

Issues with techniques like ℓ-diversity

Only practical to implement in datasets with very few variables

No computationally efficient ways of doing these; very time consuming to do by hand

- For some of the more esoteric methods, no theoretical implementations have even been described

Even if they could be implemented, in most cases achieving anything like distinct ℓ-diversity (or *t*-closeness, or *p*-sensitive k-anonymity) would completely destroy the reanalysis value of the dataset, making going to this level of effort to make data shareable rather pointless

The role of sampling

THE TWINS ARE OUT THERE

A 50% sample

Surveyed						Not Surveyed			
ID	Gender	AgeGrp	EthnicGrp	Unionize		Gender	AgeGrp	EthnicGrp	Unionize
1	M	25-30	1	**Y**		M	25-30	1	**?**
2	F	16-24	1	N		M	25-30	1	**?**
3	M	25-30	2	N		M	25-30	1	**?**
4	M	16-24	1	Y		F	16-24	1	?
5	F	31-45	1	Y		F	16-24	1	?
6	M	25-30	1	**Y**		M	16-24	2	?
7	F	16-24	1	N		F	31-45	1	?
8	F	31-45	1	Y		M	25-30	1	**?**
9	F	31-45	2	Y		M	25-30	1	**?**
10	M	25-30	2	N		M	31-45	1	?
11	M	16-24	1	Y		F	31-45	1	?
12	F	25-30	1	Y		M	25-30	2	?
13	M	25-30	1	**Y**		M	16-24	1	?
14	F	16-24	2	N		F	31-45	1	?
15	F	31-45	1	Y		F	16-24	2	?

Sampling

Creates uncertainty that any given individual is in the dataset at all

A sample unique may not be a population unique

- Still a concern…

That is, *if* an equivalence class in the dataset can be assumed to have co-equivalents (data twins) outside the dataset whose opinions or attributes are unknown, *then* attributes are not disclosed by membership in an equivalence class

This may be a reasonable assumption in cases where:

- k-anonymity is met for some value of $k > 1$
- Sample is a sufficiently small subset of the sufficiently large population it is drawn from
- There is variation across groups in the attributes being looked at

Attribute disclosure in the absence of identity disclosure ceases to be a concern in the case of a small sample drawn from a large population, given appropriate levels of variation in the attributes.

Bad examples

AND HOW I GOT INVOLVED WITH THIS STUFF IN THE FIRST PLACE

Rescuing messy data

First became seriously involved with data anonymization due to a data rescue project

Series of governmental department datasets released due to an open government mandate

Versions initially made available were unusable due to missing documentation and general incomprehensibility; we requested additional files and found they had not been sufficiently anonymized

Our contact recognized that this was a problem but had no better de-identified version of the survey, or resources for fixing it

Context: these were fairly ordinary survey datasets, with thousands of respondents, that were small samples of a large population, released without direct identifiers by people who probably thought they were taking reasonable care

They were shared in response to a mandate that hadn't been fully thought through.

Obvious problems… and not so obvious

Forward Sortation Area (first three characters of postal code) on some files – while most have populations from thousands to over 100,000, there are some with less than 50 people each. Oops?

Too many variables with small groups (allowing combinations like small ethnic group + rural PEI + age + occupation…), including surveys that asked questions about sensitive topics such as HIV status or drug use

After dealing with the obvious = got down to the k-anonymity challenge

One sample survey: 5 quasi-identifier variables of concern: age (3 categories), gender (2), geographic region (7), visible minority status (2) and Indigenous status (2) – giving 168 possible combinations

If these were distributed equally across the dataset, we would expect each equivalence class to contain about 12 cases

For most real-world variables, some groups will be much larger than others. In practice we had 21 equivalence classes with only a single member, and a total of 42 equivalence classes with less than 5 members

k-anonymity is hard

Only five quasi-identifier variables, only a few reasonable categories each...

k-anonymity is difficult to achieve in practice, and the difficulty increases as the number of quasi-identifying variables increases and the number of cases in the dataset decreases

I suspected after hand inspecting all the 1-person equivalence classes that this survey was not risky – but decided to test that.

Downloaded Census of Canada public use file, subsetting it, manipulating the variables and weighting the file to produce a dataset that matched my survey but represented the population in Canada at that time as a whole

- In effect, created an artificial census of the population my survey was drawn from

In the artificial Census dataset, the smallest equivalence class was estimated to have 370 cases, with the next smallest containing 518. So everyone had at least 369 twins

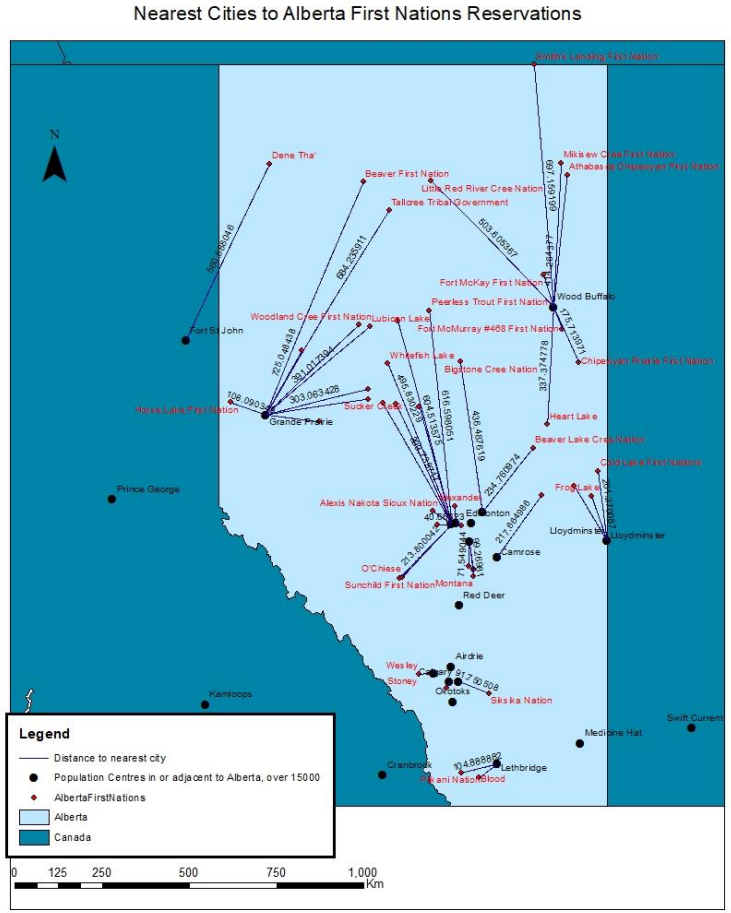
A rural service survey, GIS, and data linkage

Variables of concern: home on a reservation and distance from nearest large city

Hypothetical example:

- Participant living in Alberta on a reservation
- Response: 80 km from nearest city
- 2 possibilities only 10 km apart from one another: Samson and Ermineskin Tribe

Of over 1,000 individuals surveyed, correctly identified place of residence for 24 people (checked using FSA on file)



Hidden identifiers

"Distance from respondent's community to nearest large city" does not generally show up on lists of possible identifiers or quasi-identifiers to check for.

Variables that are not obviously risky may be used in combination to derive other quasi-identifiers and data linkages can make unlikely leaps – I haven't even touched on more complex things being done with fuzzy logic and AI …

What do we do? Burn it all down?

Final observations and recommendations

Guaranteeing that data has been reasonably anonymized is difficult, and the difficulty increases exponentially with the number of potentially identifying variables present.

k-anonymity can be *calculated* easily using standard statistical software. *Achieving* k-anonymity can require a great deal of data modification.

A small sample of a large population (with no way to distinguish participants from non-participants) is much less risky. A dataset that is a complete or large sample of a known population is very difficult to deidentify unless the number of variables is trivial.

In general, take a harm avoidance approach

- Don't share the survey with HIV status unless you're really sure you know what you are doing
- Don't share data gathered from a known group (hocky referees in Ontario, members of a Facebook group for leukaemia survivors, etc) unless consent waives confidentiality
- Consider the dataset as a whole – number of identifiers, anything geography, anything that might show up somewhere else (including non-demographics), overall potential for harm of the questions

If after this you're dubious about whether the dataset's safe… don't share openly.

Sources and further reading

Ayala-Rivera, V., McDonagh, P., Cerqueus, T. and Murphy, L. (2014) 'A systematic comparison and evaluation of k-anonymization algorithms for practitioners', Transactions on data privacy, 7(3), pp.337-370.

British Medical Journal (2010) Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. BMJ 2010; 340.

Information and Privacy Commission of Ontario. De-identification Guidelines for Structured Data.

Portage Covid-19 Working Group (2020) De-identification Guidance.

Samarati, P. and Sweeney, L. (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Smith, D. (2020) Re-identification in the Absence of Common Variables for Matching. International Statistical Review, Volume 88, Issue2.

Thompson, K and Sullivan, C. (2020) Mathematics, risk and messy survey data, IASSIST Quarterly 44 (4).

U.K. Anonymization Network. The Anonymization Decision-Making Framework, 2nd ed.

Artwork created with the help of Artificial Intelligence using NightCafe Creator with the DALL·E 2 algorithm.