Reducing Risk: An Introduction to Survey Data Anonymization

Creating safe(r) shareable data

Western Libraries

portage

SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

Background and key concepts

Identifiers, quasi-identifiers, risk

That data rescue project

- First became seriously involved with data anonymization due to a data rescue project
- A bunch of survey files from Health Canada released under Open Government mandate
- Group of Ontario data library people worked to create cleaner and more user-friendly versions of some of the ones found on Canada’s open government portal
- Bright idea – ask Health Canada if they had data from some additional surveys that were mentioned but not released
- They sent me some files…

Oh dear.

- Census geography, parts of postal codes, telephone area codes, you name it.

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

The first step in data anonymization is always to locate and remove or mask all direct identifiers

But wait! Your job isn't done!

Quasi-identifiers

- Characteristics relating to individuals that could be linked with other data sources to violate the confidentiality of individuals
- A variable should be considered a quasi-identifier if an attacker could plausibly match that variable to information from another source to determine the identity of an individual
- Some variables may be used in combination to derive quasi-identifiers, e.g. community size (at first glance not particularly identifying) could be combined with a broader geographic grouping to infer location more precisely

Hidden 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

Risk – a technical definition

- 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
- 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*.

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 better wifi”: not identifying
  - “And when I spoke to my colleagues at the plant about organizing a union …” possibly identifying

Assessing and dealing with risk: statistical disclosure risk assessment

Heuristics and an introduction to k-anonymity

Western Libraries

portage

SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

Assessing quasi-identifiers

- Quasi-identifying variables containing groups with small numbers of respondents (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

Common sense

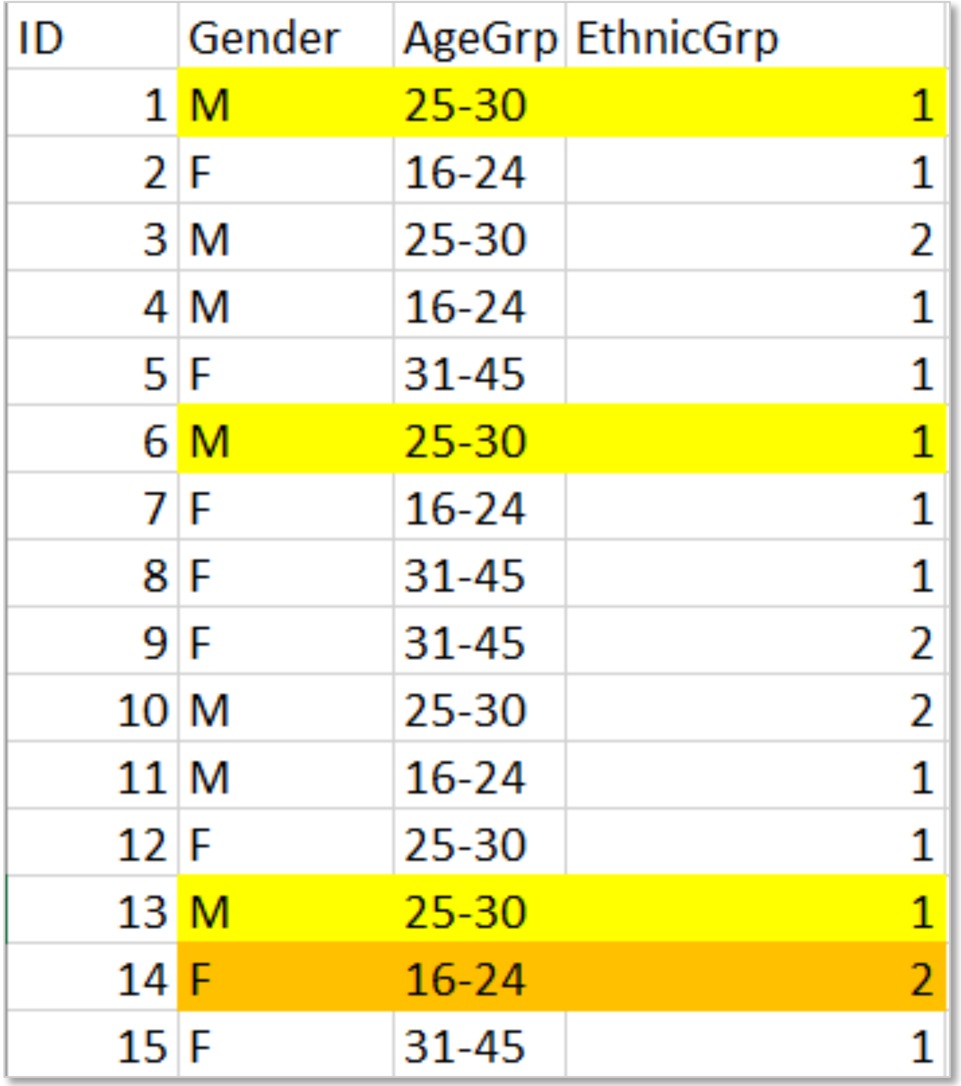
- Look at the demographic variables in the dataset and consider describing an individual to a friend using only the values of those variables. Is there any likelihood that the person would be recognizable?
- "I'm thinking of a person living in Toronto who is female, married, has a University degree, 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…
- Also, consider unusual combinations of variables – let's say someone belongs to the under-16 age group and also responded that they were married.

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 identifying variables
- That is, a record cannot be distinguished from K-1 other records in its equivalence class.
- K is a number set by the researcher; three and five are both commonly used
- Values higher than fifteen are rarely used, according to one article I found. In practice I have not seen a value higher than five referenced and this is the number most frequently referred to in the literature.

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 'under 16' age group who responded 'married' might have their response to the marriage question deleted as an alternative to further recoding the otherwise non-risky variables of AgeGroup or 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 < 3, 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.

Issues with k-anonymity

Adding ℓ-diversity and other lettered concepts into the mix

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 record in the anonymized data, will be indistinguishable from other k-1 records in the same dataset

However…

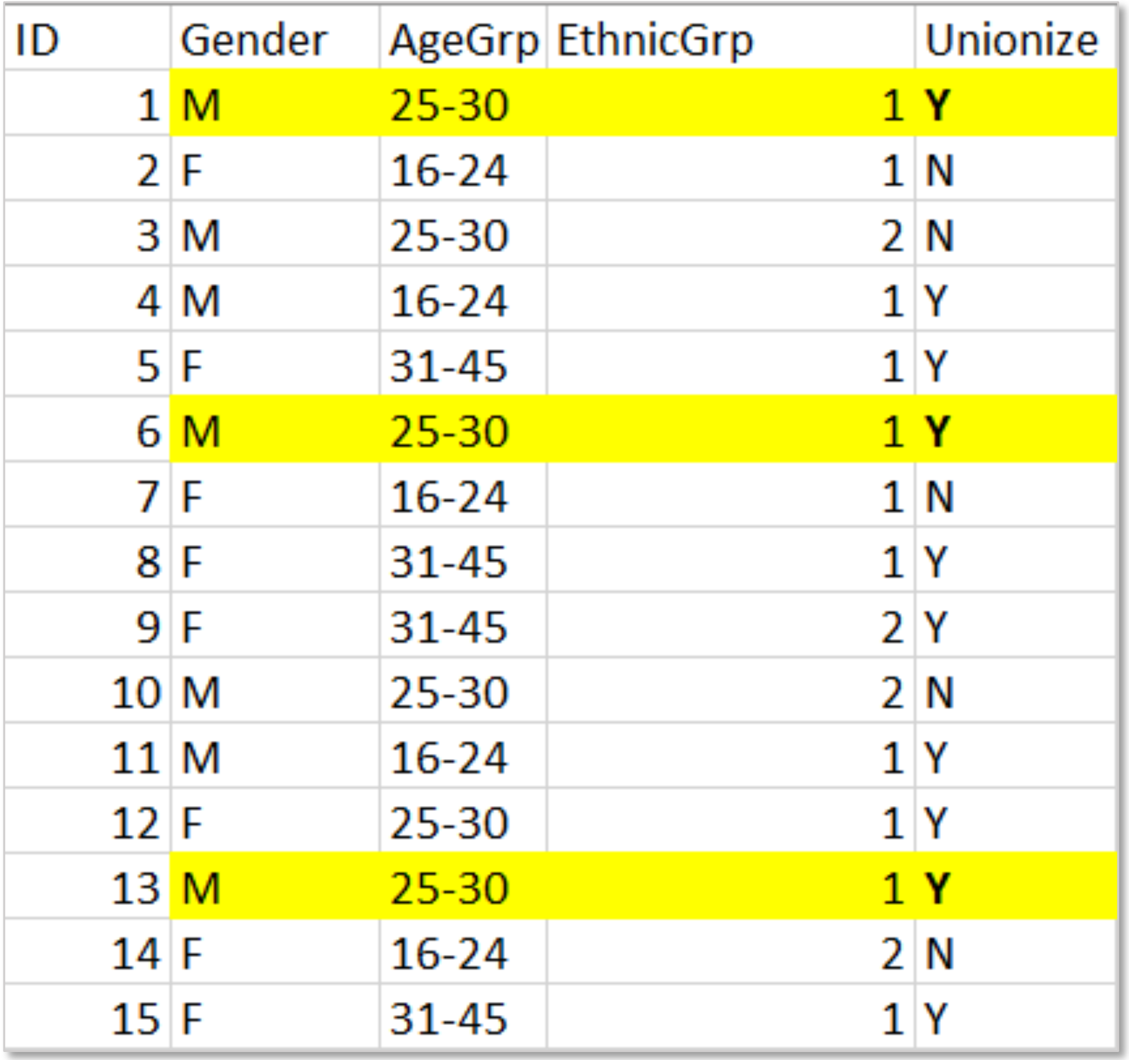
- Survey respondents are not generally told that no one will know which line of the data file holds their survey responses. They are told their answers to survey questions will be kept confidential.

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

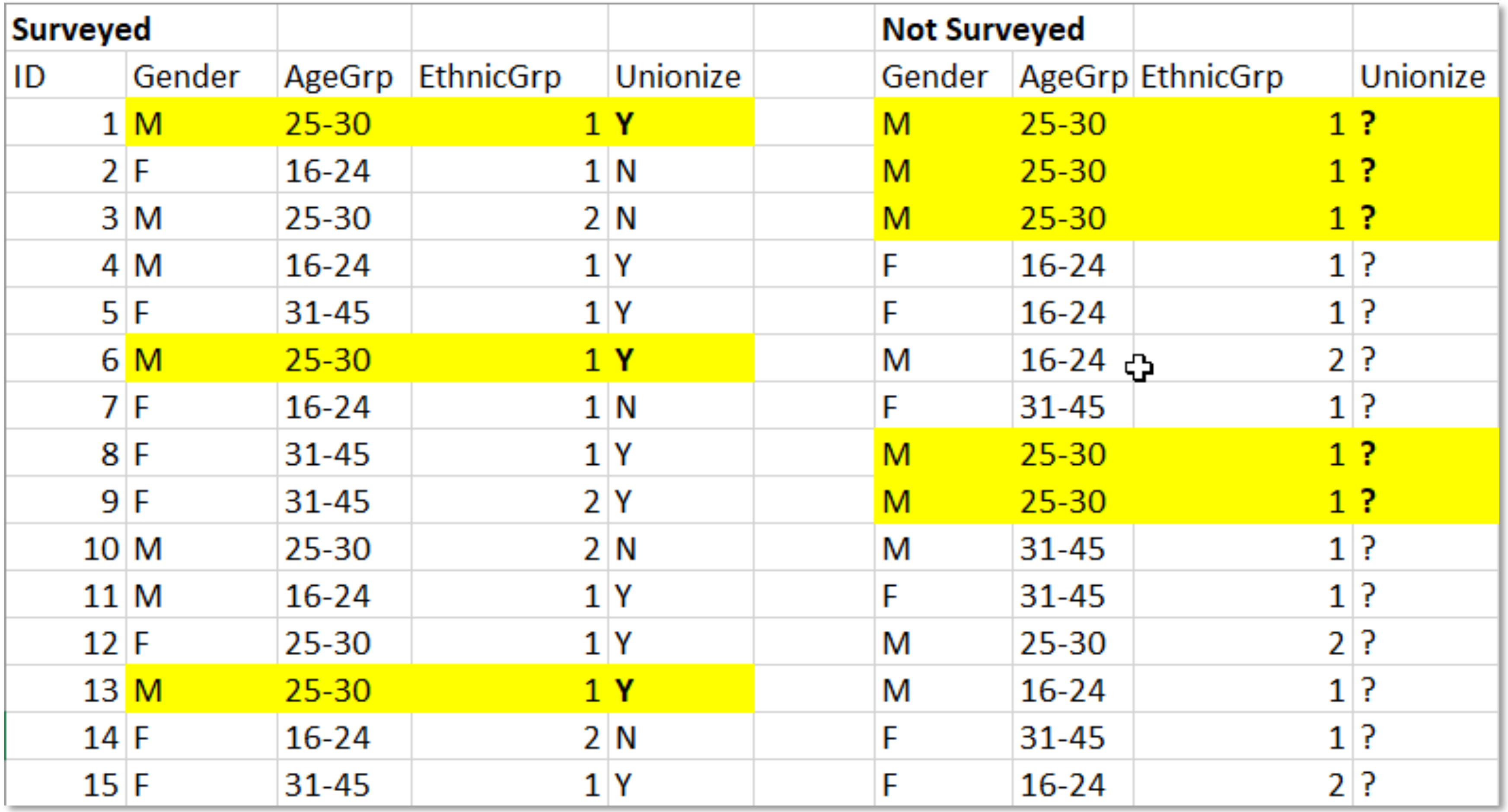
- Extensions of *k*-anonymity, including *p*-anonymity and ℓ-diversity, have been proposed to deal with attribute disclosure; they all involve rules around what values the attributes within an equivalence class should have
- Example: one of the simpler variants, called distinct ℓ-diversity
  - A dataset 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; for large datasets, far too time consuming to be done 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 ℓ-diversity (or *t*-closeness, or *p*-diversity) 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

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 is a reasonable assumption in cases where:
  - k-anonymity is met for k >=5
  - Sample is a small subset of the population it is drawn from
- 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

Other strategies – increasing the randomization in the data

- Perturbation / random noise
  - Adding or subtracting a random number following some distribution from numeric variables.
  - Maintains overall distribution of data but changes the error term in the data in a way that is opaque to the researcher. Not suitable for categorical variables.
- 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. May change bivariate / multivariate distributions.
- Some statisticians love these things, but they have not gained traction among researchers
  - Doing them in a sensibly data-utility preserving way is complex and may required that the data analyst be able to compensate for bias and distortion

Case study

Health Canada drug use survey

That data rescue project

- National Anti-Drug Strategy Survey Series – a set of surveys of adolescents asking questions about drug use
- 1502 respondents, 339 variables including limited demographics
- Five quasi-identifier variables of concern: age (3 categories), sex (2), geographic region (7), visible minority status (2) and aboriginal status (2)
  - 126 Possible equivalence classes (not 168 because visible minority and aboriginal are mutually exclusive as defined (ask Statistics Canada))
- 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 categories each
- Fairly large dataset
- We were not able to produce a dataset that satisfies k-anonymity, let alone any more stringent criteria such as l-diversity, while retaining all five variables
- We were able to achieve k-anonymity by deleting the region variable; on the remaining four variables there were no equivalence classes smaller than 5.
- 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

The role of sampling, redux

- How risky would it have been to retain the region variable? Were our sample unique cases (the 21 equivalence classes with only a single member) also population uniques?
- Checked by downloading a Census of Canada public use file, subsetting it and manipulating the variables, and weighting the file to produce a dataset that matched my survey but represented the population aged 13-15 in Canada at that time as a whole
  - In effect, created an artificial census of the population my survey was drawn from
- In the Census dataset, the smallest equivalence class was estimated to have 370 cases, with the next smallest containing 518, and the remaining 214 equivalence classes being considerably larger
- Each sample unique in the drug use survey is estimatd to have a minimum of 369 data twins in the general population – k-anonymity overestimated reidentification risk by a factor of 370!

Removing the region variable

- Given the sensitivity of this survey, removing the region variable probably made sense but leaving it in would be defensible *given* that I checked population risk using the census
- Most data curators are not going to go to the effort of using the census to check population risk, and results will not always be as dramatic as the ones in this example
- Most of the literature on data anonymization neglects the massive effect that sampling has on reducing risk – anonymization software also tends to ignore this factor

In practice

- For the data curator, it makes sense to look at k-anonymity as a way of safeguarding identity disclosure in sensitive data, while relying on the sampling effect to deal with attribute disclosure in the case of a small sample drawn from a large population
  - Complete or near-complete samples of smaller, defined populations (e.g. a single workplace) are inherently risky and the curator may want to consider other options for sharing
- Checking k-anonymity is not difficult using standard statistical software packages
- To deidentify a dataset using common statistical software, the following steps may prove helpful:
  - Identify and remove or mask direct identifiers, identify quasi-identifiers of concern
  - Use frequencies and bivariate tables to identify small groups and iteratively create larger groupings using data reduction
  - Check k-anonymity using code presented, inspect small equivalence classes, use the common-sense approach to determine if these are truly risky. Where uncertain, regroup variables or suppress.
- This is *probably* adequate for a dataset that is a small sample and is not inherently high risk

Automating it

Software solutions

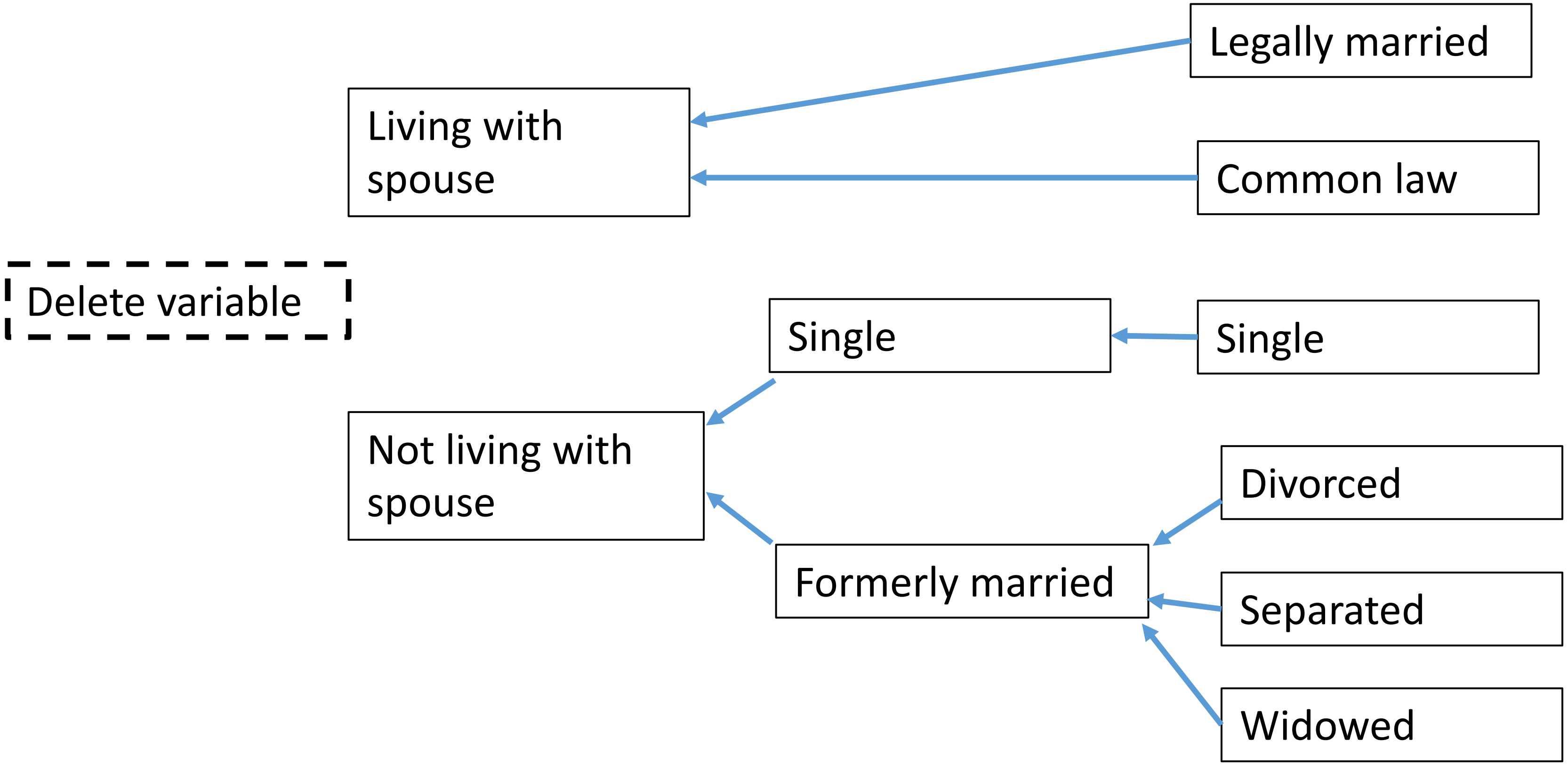
Amnesia and SDCMicro

- While working on the initial deidentification project and later while contributing to some documents for a working group, I tested several anonymization packages that I found recommended on various lists.
- The two that seemed most functional (although still with some shortcomings particularly in documentation) were SDCMicro, an R package with a graphical interface, and Amnesia
- Both will check data for k-anonymity given a list of quasi-identifiers, and provide tools for dealing with direct identifiers
- I preferred SDCMicro for usability and because it correctly handled missing data (Amnesia doesn't allow missing value specification) but both have similar approaches to automatically adjusting quasi-identifiers

Anonymization hierarchies

- These tools take a hierarchy approach to automatically deidentifying data with quasi-identifiers.
- This basically means that the user needs to pre-define possible generalizations for the quasi-identifiers in the dataset, and the program will search for possible solutions and recommend a set of the generalizations to use
- For datasets with a large number of quasi-identifiers, or cases where several datasets with similar quasi-identifiers need to be deidentified, this might be a useful approach. In the data I have worked with I found it as easy to do by hand.

Possible hierarchy for the variable “Marital Status”



Final observations

- Guaranteeing that data has been completely 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 or suppression, though the role of sampling somewhat mitigates this
- Software aimed at the general academic survey researcher should not assume special knowledge in the field of data de-identification. I didn’t find any packages I would unreservedly recommend, out of the 6 packages I tried

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. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *BMJ* 2010; 340 (Available at https://doi.org/10.1136/bmj.c181)

Domingo-Ferrer, J. and Torra, V. (2008) ‘A critique of k-anonymity and some of its enhancements’, In Third International Conference on Availability, Reliability and Security. IEEE.

Elliot, M., Mackey, E., O'Hara, K. and Tudor, C. (2016) *The Anonymisation Decision-Making Framework*, Manchester: UKAN. (Available at https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf)

Samarati, P. and Sweeney, L. (1998) ‘Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression’, (available at https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf)

- Perhaps asking people to share experiences or thoughts on ethical issues with data sharing