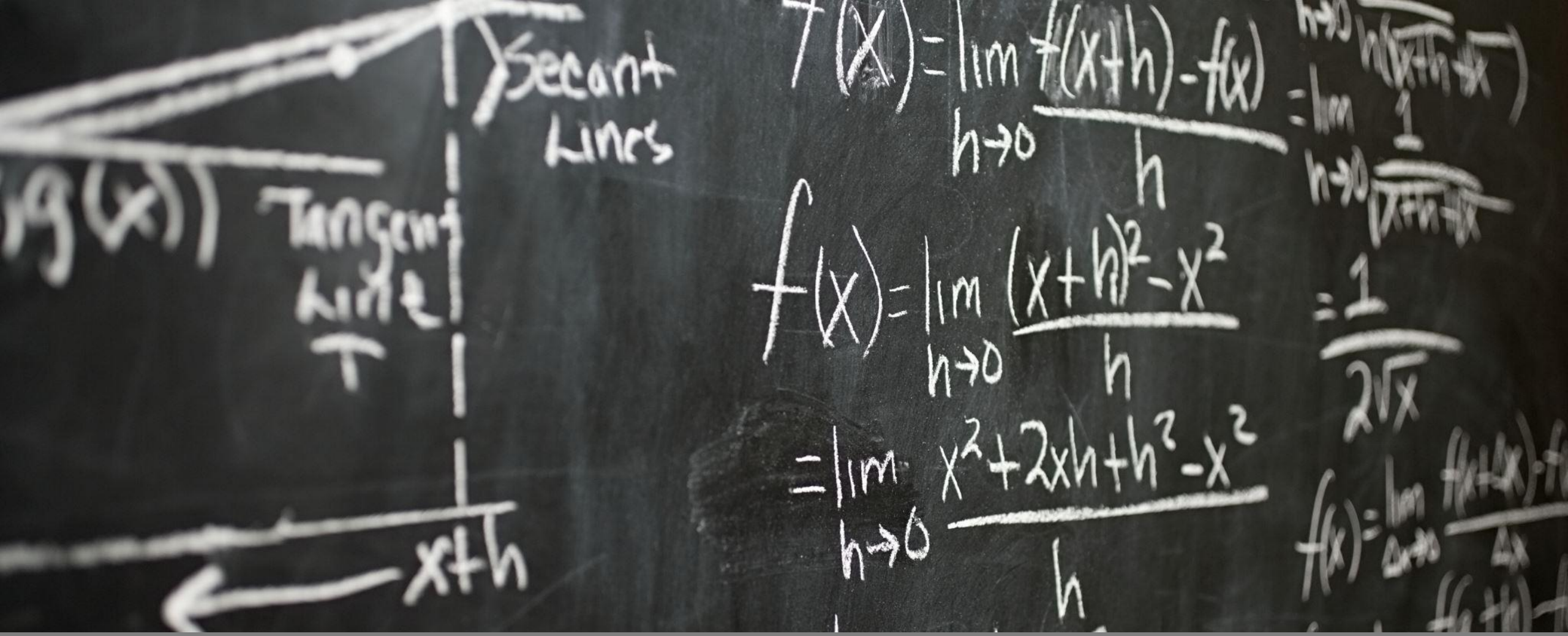
3-3-2022

The Mathematics of Risk: An introduction to guaranteed data de-identification

Kristi Thompson
Western University



The Mathematics of Risk

An introduction to guaranteed data de-identification

?

Going about things backwards?

Dr. Khaled El Emam, Children's Hospital of Eastern Ontario, "Privacy enhancing technologies to enable the sharing of health data for secondary purposes."

- How do you create safe, shareable versions of data that can't be sufficiently de-identified?

Vance Lockton, Information and Privacy Commissioner's Office, Ontario, "What is 'reasonable'? Exploring when a de-identification process is sufficiently protective."

- How do you decide when data has been reasonably deidentified?

Kristi Thompson, Western University, London Ontario, "The mathematics of risk: an introduction to guaranteed data de-identification."

- What is this de-identification thing anyway and why is it so hard?

Some context

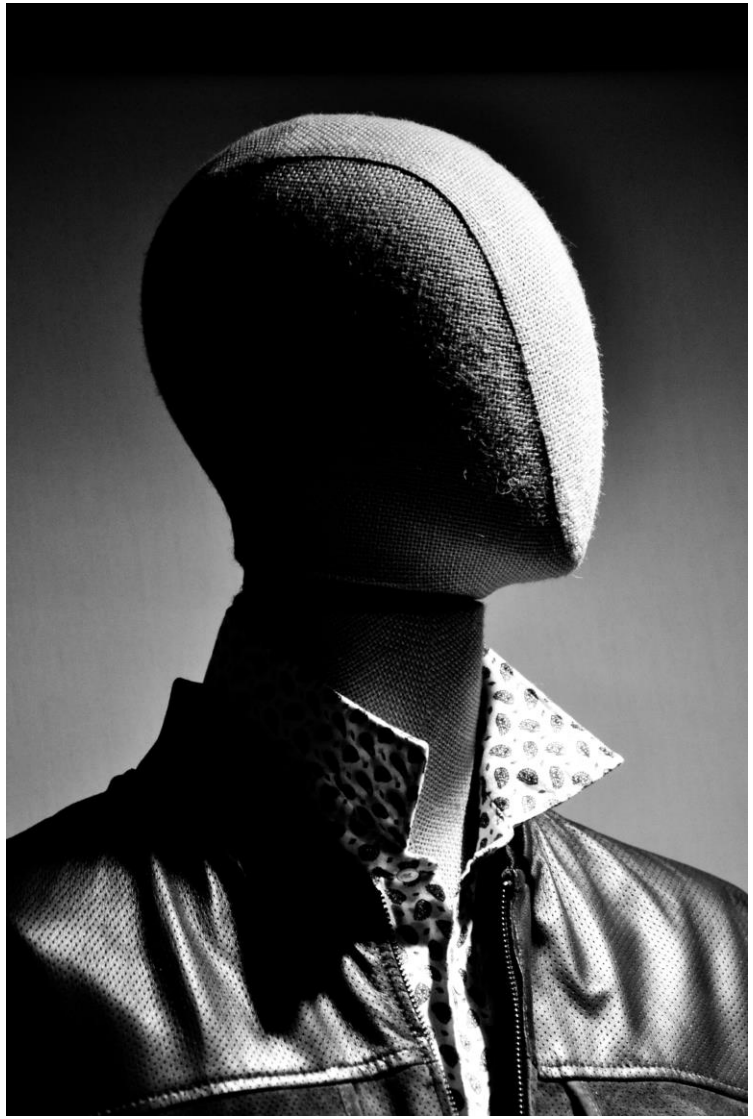
"Grant recipients are required to deposit into a digital repository all digital research data, metadata and code that directly support the research conclusions in journal publications and pre-prints that arise from agency-supported research."

- From the Tri-Agency Research Data Management Policy

This 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.

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

- Are researchers expected to understand when data deidentification is needed to comply with this policy? 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

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*.

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…
- **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.

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

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 < 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.

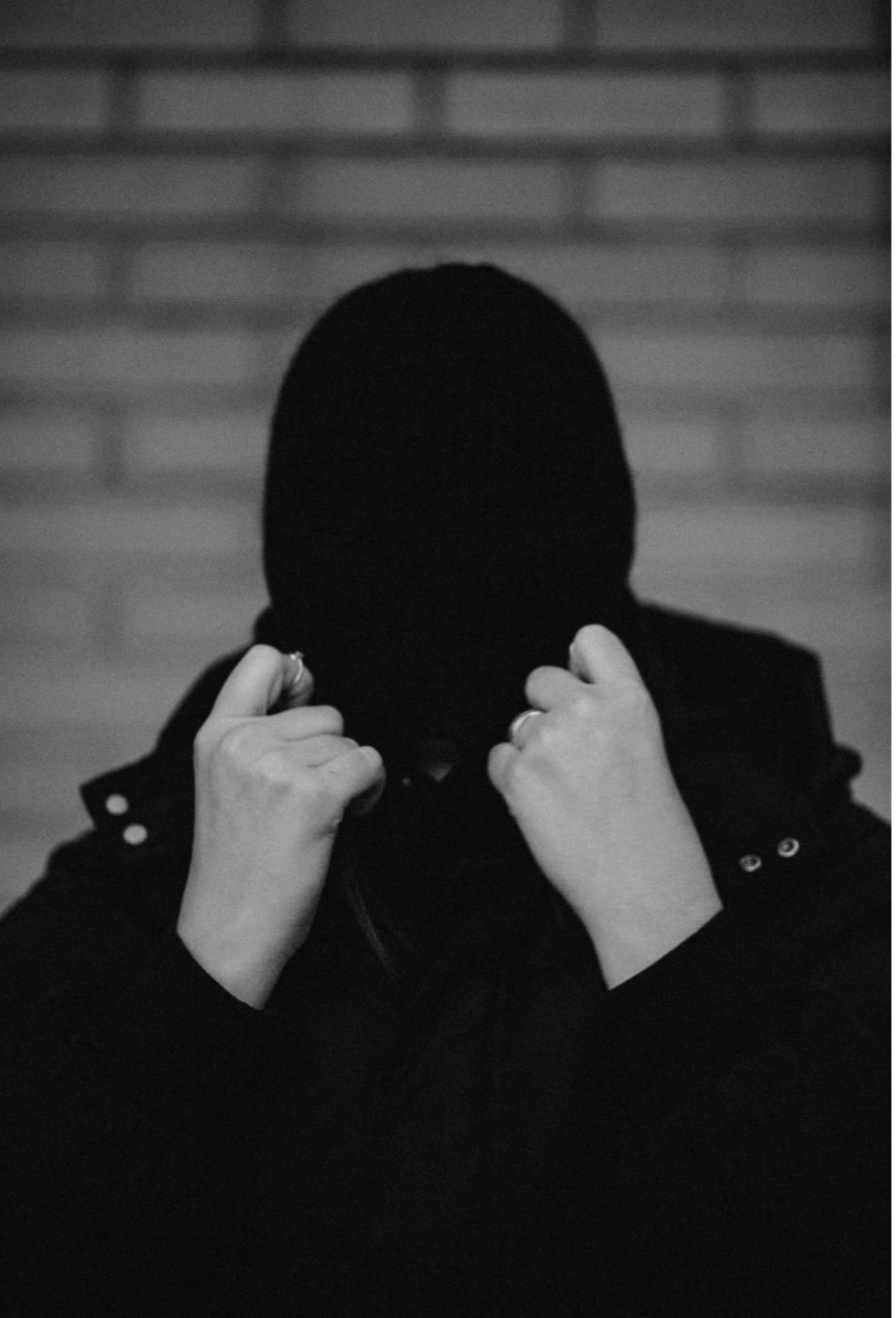
~~Guaranteed~~ data anonymization

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

It guarantees that every record in the anonymized data will be indistinguishable from k-1 other records in the same dataset.

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*-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; far too time consuming to be done by hand

- For some of the more esoteric methods, no theoretical implementations have even been described
- It's been demonstrated that even in relatively simple cases (such as ℓ-diversity with few attributes) automatedly solving for optimal data utility while protecting privacy is NP hard – meaning, essentially, that the time taken to run such an algorithm increases exponentially with the size of the dataset

Even if they could be implemented, in most cases achieving anything like distinct 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
- There is variation 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.

Practical examples: sampling and geography

And how I got involved with this stuff in the first place

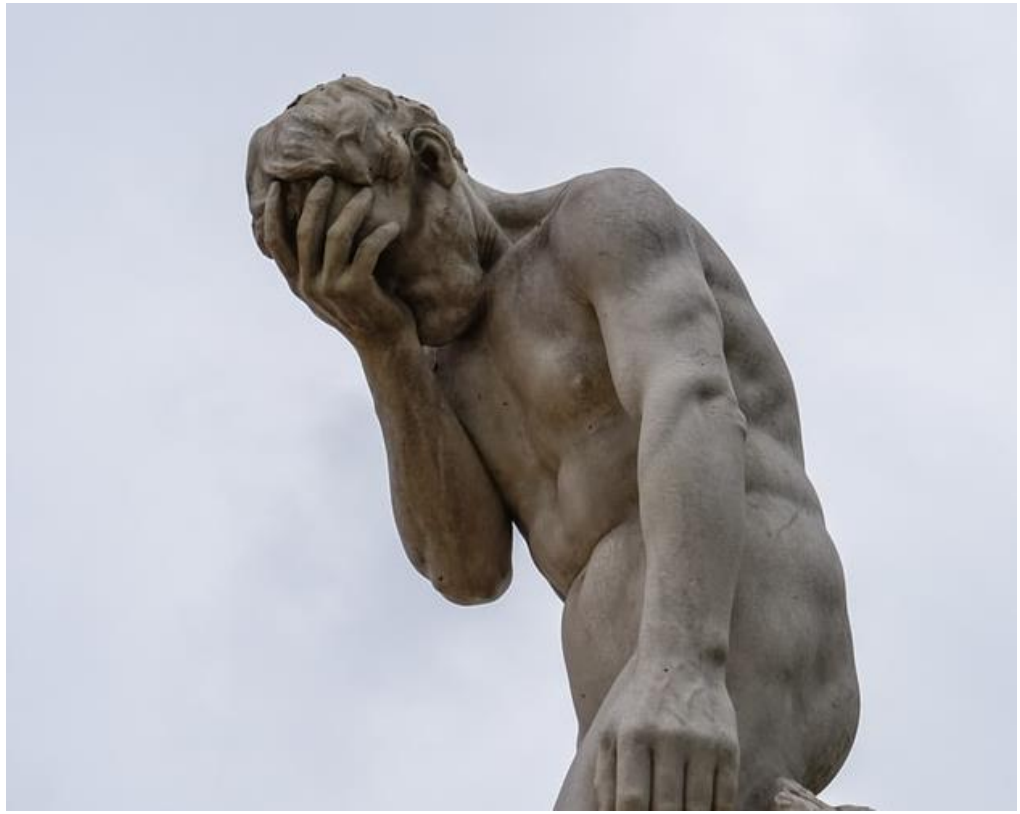
Rescuing messy data

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

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

Versions initially made available were unusable due to missing documentation and general incomprehensibility; documented versions made available on request had not been anonymized.

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



The first test survey

Survey of adolescents asking about an ad campaign

~1500 respondents, limited demographics, various non-identifying variables

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 status are mutually exclusive as defined (...ask them))

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

* Aboriginal is the term used in these surveys, currently Indigenous is the preferred term

k-anonymity is hard

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

Fairly large dataset

Was not able to produce a dataset that satisfied k-anonymity, let alone any more stringent criteria such as l-diversity, while retaining all five variables

Was 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 the 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, 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 artificial 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 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!

Second test survey

Additional survey with similar demographic variables, plus additional data associated with participant location

In addition to checking k-anonymity based on demographic variables (in this survey, limited) wanted to look at the geographic variables

Service survey of a population living in small and remote communities. Did not have exact location, community names, or anything obvious like that except province…

But **did** identify some respondents as being located on a reservation, and also had a variable giving approximate distance to nearest major city

Original un-de-identified dataset also had partial postal codes that could be used to check guesses

Penetration testing and data linkage

Means of assessing resistance of de-identified dataset to reidentification of survey participants or their attributes

Remember: quasi-identifiers contain information that can be matched to persistent information that an attacker may reasonably have access to

From publicly available information, a data intruder can easily construct a table of reservations by province and their distance from the nearest city

For each participant, their province of residence and distance from the nearest city can be compared to the entries in the table of reservations by province and distance to the nearest city

Use of data linkage to construct lists of candidate locations for survey participants

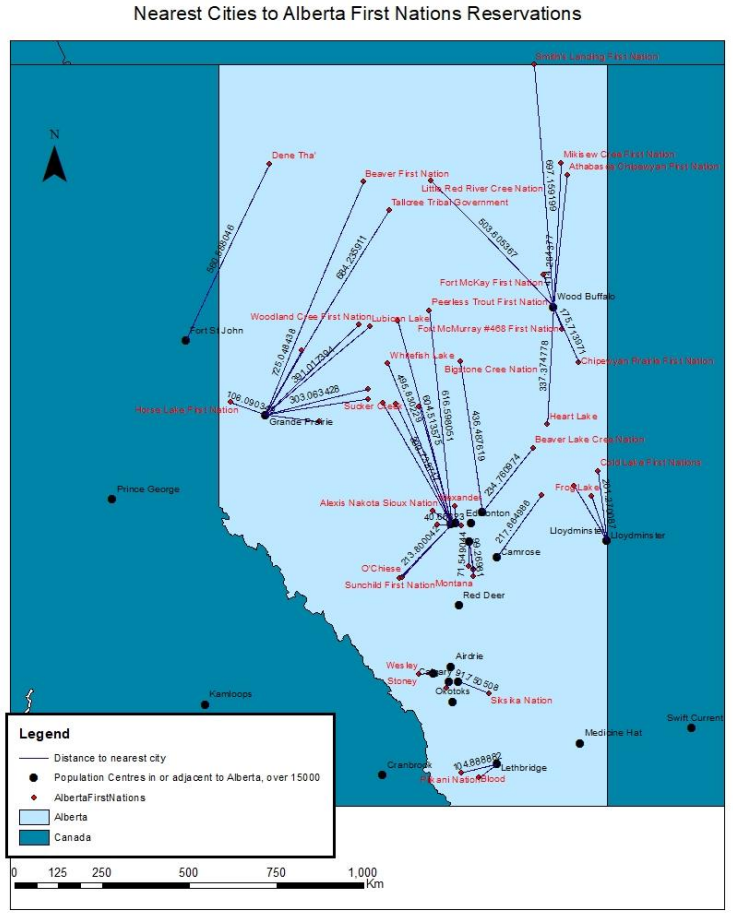
Hypothetical example:

- 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, a single location for their potential place of residence was found for 98

Of the 98, the (suppressed) value for forward sortation area (first three digits of postal code) was correct for 24 cases (~25% of guesses)

Accuracy could be improved with access to more specialized GIS tools



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

- Community name, yes.

Variables that are not obviously risky may be used in combination to derive other quasi-identifiers

- So burn it all down?

Anything relating to geography needs to be considered. By extension, there might be similar variables that can indirectly identify other groups, such as clubs, organizations, or employers.

Automation

ALGORITHMIC OPTIONS

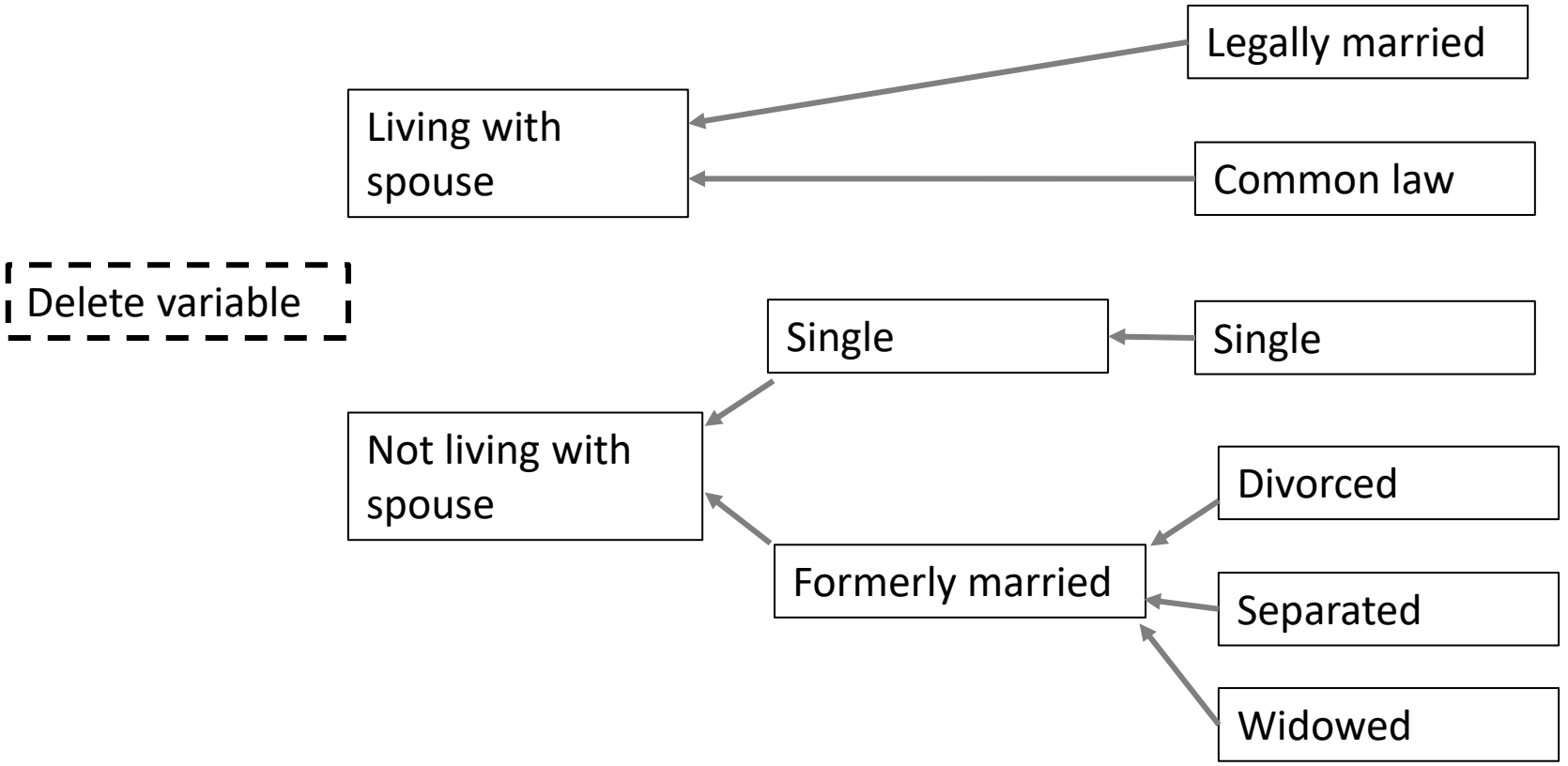
Anonymization hierarchies

Software tools for de-identifying quantitative data that I have investigated take a hierarchy approach to automatically deidentifying data.

This basically means that the user needs to pre-define possible generalizations for the quasi-identifiers in the dataset, and the program searches for possible solutions and recommends a set of the generalizations to use.

For datasets with many quasi-identifiers, or cases where several datasets with similar quasi-identifiers need to be deidentified, this might be a useful approach… if a tool can be found that actually does the job.

Possible hierarchy for the variable "Marital Status"



Tools implementing these approaches

While working on the initial deidentification project and later while contributing to some documents for a working group, I tested several free / open 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 had usability issues, were not adequately documented, and Amnesia didn't seem to handle missing values correctly. Online reviews suggested that all the packages had issues handling large or complex datasets.

Commercial software and for-fee services exist. I have not had the opportunity to try any.

Final observations

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 or suppression, though the role of sampling somewhat mitigates this. A dataset that is a complete sample of a small, known population is very difficult to deidentify unless the number of demographic and attribute variables is trivial.

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 free packages I would really recommend, out of the 6 packages I tried.

The new Tri-Agency policy on Research Data Management mandates open sharing of research data where possible. Most of the academic social science researchers I know do not have the knowledge to assess their data for anonymization. The data curators who administer university-based data repositories are similarly unequipped. We desperately need better supports.

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. 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

Portage Covid-19 Working Group. 'De-identification Guidance', available at https://zenodo.org/record/4270551#.Ygvklt_MKM8

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

Thompson, K and Sullivan, C. (2020) 'Mathematics, risk and messy survey data', IASSIST Quarterly 44 (4), available at https://iassistquarterly.com/index.php/iassist/article/download/979/961