Mathematics, risk, and messy survey data

Kristi Thompson and Carolyn Sullivan[1]

Abstract

Research funder mandates, such as those from the U.S. National Science Foundation (2011), the Canadian Tri-Agency (Social Sciences and Humanities Research Council, 2018), and the UK Economic and Social Research Council (2018) now often include requirements for data curation, including where possible data sharing in an approved archive. Data curators need to be prepared for the potential that researchers who have not previously shared data will need assistance with cleaning and depositing datasets so that they can meet these requirements and maintain funding. Data de-identification or anonymization is a major ethical concern in cases where survey data is to be shared, and one which data professionals may find themselves ill-equipped to deal with. This article is intended to provide an accessible and practical introduction to the theory and concepts behind data anonymization and risk assessment, will describe a couple of case studies that demonstrate how these methods were carried out on actual datasets requiring anonymization, and discuss some of the difficulties encountered. Much of the literature dealing with statistical risk assessment of anonymized data is abstract and aimed at computer scientists and mathematicians, while material aimed at practitioners often does not consider more recent developments in the theory of data anonymization. We hope that this article will help bridge this gap.

Keywords

anonymization, deidentification, sensitive data, survey data

Introduction

As a result of an open government mandate in 2014 (Government of Canada, 2014), many datasets and surveys were released on Canada's government open data portal, giving researchers access to a trove of previously unavailable survey data. Among these datasets were a set of surveys that had been conducted by firms on behalf of Health Canada. These surveys came to the attention of a group of Canadian data librarians in part because the files were in several cases released in formats that were difficult to use and without adequate documentation. In other cases, documents were released that discussed surveys that had not been made available. The efforts of this group to track down documentation and datasets and obtain or build easier-to-use survey files have been described in a previous article (Thompson, 2018). As a result of these efforts, one of the co-authors of this article was entrusted with a small collection of Health Canada datasets, which needed additional work beyond the group's now standard practice of documenting and formatting. Among the issues were that the datasets had not been fully assessed for anonymization, though obvious identifiers like name and address had been removed. The first coauthor, a data librarian with a computer science background, spent some time familiarizing herself with the theory behind data anonymization and statistical disclosure risk assessment in order to deal with this unexpected data issue. The second coauthor, a Master of Library Science student who also has a computer science background, became involved later when work on the data became a part of her internship in data librarianship with the first co-author. This paper will provide a background to the field of data anonymization and explain the work we did to ensure that two particularly difficult datasets had been anonymized. Based on our experiences, we will describe a set of processes for reviewing messy survey data to make sure it has been properly anonymized.

Data anonymization background

The first step in data anonymization is the removal of all direct personal identifiers - data elements that can be directly linked to a specific individual such as names, telephone numbers, social media identifiers, and so on. This step is obvious and inexperienced data curators may assume it is sufficient. However, demographic variables can also pose risk. A common example might be occupation and geography; if the dataset includes name of town and occupation, then the only doctor in a small town is at risk of being identified. Add additional variables such as ethnicity, country of origin for immigrants, gender, and family structure, and even doctors in larger cities might be at risk of reidentification. These variables - persistent demographic characteristics of people that might be used to discover their identities - are known as quasi-identifiers. The problem of reidentification with quasi-identifiers becomes more acute when you consider not just the unlikely possibility of the hypothetical doctors' neighbours reading through the dataset and recognizing them, but the unfortunately plausible possibility that an intruder might use external information from directories and other public sources to attempt to reidentify people maliciously, for fun, or for profit.

A variable should only be considered a quasi-identifier if an intruder could plausibly match that variable to information from another source. Some variables may be used to derive other quasi-identifiers; for example, community size could be combined with a broader geographic grouping to guess the precise community someone lives in. Remaining variables in the dataset are non-identifying variables and will include opinions and ratings, temporary measures such as recent food consumption or exercise, and other research questions. Risk is created when there is the potential for an intruder to link external information to identifiers or quasi-identifiers in a dataset to gain additional information about individuals.

The question then becomes, how do you decide whether there are people in a dataset who are at risk of being reidentified, and how do you keep this from happening? In the example above, an obvious approach might be to remove the quasi-identifiers of 'town' and 'occupation' from the dataset. But there might still be some unusual combination of ethnicity, family structure, age and other demographic variables that could lead to the reidentification of an individual. Removing all quasi-identifiers would remove risk from the dataset but is usually overkill; a variable such as gender or age on its own does not pose significant risk. And demographic variables are of great utility to researchers. Every variable removed from the dataset decreases the dataset's research value.

Rather than removing variables completely, a common practice is to group values into broad categories; a variable such as occupation might be grouped into categories of professional and non-professional occupations or according to some other scheme, age might be grouped into 10 year categories, and similar measures may be taken with other variables. But the question remains: how can we be sure we have done enough to ensure that participants remain anonymous?

Introduction to k-anonymity

A common approach is k-anonymity, one type of what is called data analytic risk assessment (DARA) or statistical disclosure risk assessment (Eliot, McKay et al, 2015). K-anonymity was first developed by computer scientists Sweeney and Samarati (1998) and has formed the basis of formal data anonymization efforts since then. Ayala-Rivera, McDonagh et al (2014) describe it as a 'fundamental principle of privacy' and 'widely discussed and adopted in a variety of domains'. Simi et al call it the 'primary model proposed for microdata anonymization' and 'the base from which further expansions have been developed.' The concept behind k-anonymity is relatively straightforward: 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, typically 5. That is, a record cannot be distinguished from k-1 other records. A minimum of 3 is commonly suggested for k; in practice a value of 5 is often used. According to El Emam and Dankar (2008) data custodians should 'select a value of *k* commensurate with the re-identification probability they are willing to tolerate—a *threshold risk*.' They also note that 'it is uncommon for data custodians to use values of *k* above 5 .'

An example may help illustrate this concept. Imagine your survey has four demographic variables: marital status, age group, gender, and ethnic group. If an individual in the dataset is white, married, over 65, and female, then for the data to have k-anonymity with k=5, there must be at least four other individuals in the dataset with the same set of characteristics. This also must be true for every other individual in the dataset; each person must have at least four data twins. Even if an intruder knew that an individual was in the dataset and was able to match their characteristics against the data, they would not be able to tell which of the five cases was the target individual.

A set of data twins, or cases with the same values on all potentially identifying variables, is called an equivalence class. An individual in the dataset who does not have any data twins - an equivalence class of one - is called a sample unique. This individual is at risk of being re-identified. If the dataset is a complete sample of a small population (for example, employees at a particular company) then this sample unique will also be a population unique, and an intruder looking at this dataset will be able to definitively identify this person. Even if the dataset is not a complete sample, it is still possible that this person may be a population unique or have some combination of rare characteristics that makes it easy to narrow down their identity to a small number of candidates.

K-anonymity is not always sufficient

K-anonymity is a useful technique for limiting the possibility of exposure of the identity of the individuals in a data set. However, it may not always be sufficient to prohibit *attribute* disclosure, as many articles have noted. Consider the above dataset, hypothetically a complete sample of all employees at a particular company. By grouping and categorizing variables, we have achieved k-anonymity with a k of 5 - the smallest equivalence class in the dataset contains at least five respondents. None of the individuals in the dataset can be identified with any certainty. However, imagine that all the respondents in an equivalence class answered a question the same way - for example, all the employees at our company with a specific combination of characteristics answered 'yes' to a question about organizing a union. If a manager at this company looked at the data, that manager would now know an attribute (interest in joining a union) about everyone in that equivalence class. K-anonymity is not sufficient to prevent attribute disclosure. Respondents to surveys are generally told that their responses will be kept confidential, not merely that no one will know which line of data contains their specific answers; a k-anonymous dataset may not fulfill that promise.

Variations on k-anonymity, such as p-anonymity and l-diversity, have been developed in an effort to deal with the attribute disclosure issue. However, as Domingo-Ferrer and Torra (2008) note in their review of k-anonymity and its variants, 'neither k-anonymity nor its enhancements … are entirely successful in ensuring that no privacy leakage occurs while keeping a reasonable data utility level.' A brief explanation of one of the variants of l-diversity will serve to illustrate this problem.

A data set is said to satisfy distinct l-diversity if, for each group of records sharing a combination of demographic attributes (an equivalence class), there are at least *l* different values for each confidential variable. In our example workplace dataset, every group of data twins would need to include both yes and no responses to the union question. However, the effort needed to check for and recode to ensure this by hand, assuming that, as in many datasets, there are dozens of responses that need to be kept confidential, is daunting. In addition, one can easily imagine a scenario where nearly every attribute needed to be manipulated, grouped or partially suppressed in some way, or alternatively every equivalence class needed to be enlarged. As we will see later in this paper, a relatively small set of quasi-identifying variables can easily lead to a very large number of potential equivalence classes in a dataset. Multiply that number by the set of attribute responses that would need to be considered and you will have a sense of the scope of the problem. At the end of all that manipulation, the resulting dataset would probably have very limited analytic value. The manipulation would have the result of destroying close relationships between the quasi-identifying demographic variables and the remaining variables in the dataset and determining accurately the correlation between demographic variables and attributes is a key part of research.

K-anonymity is not always necessary

Until now we have been considering an example dataset which surveys an entire population - a set of workers at a particular place of employment. It is very difficult to ensure confidentiality of responses in datasets of this nature. However, many datasets do not survey entire populations but are instead sample surveys. Imagine that only one in 10 of the workers were surveyed. Even if a worker belonged to an equivalence class of one, it would not be clear from the dataset whether that worker might have 'data twins' outside the dataset - a sample unique might or might not be a population unique, although someone with perfect knowledge of the population being sampled might still be able to determine that individual's identity in the case that they were a true population unique. Sampling can also protect against attribute disclosure, since members of an equivalence class are likely to have data twins outside the dataset whose potential responses to confidential questions are unknown. According to Eliot, McKay et al (2015), 'sampling is one of the most powerful tools in the toolbox. The key point is that it creates uncertainty that any given population unit is even in the data at all.' The larger the population from which a sample is drawn, the less likely an intruder is to have perfect knowledge of the population, and the less likely it is that a sample unique will also be a population unique. We argue that attribute disclosure in the absence of identity disclosure is not a genuine concern in the case of a sample being drawn from a large population. If an equivalence class in the dataset can be assumed to have many co-equivalents in the general population being sampled whose non-identifying attributes or opinions are unknown, then membership in an equivalence class cannot be said to reveal confidential attributes or opinions[2].

National Anti-Drug Strategy Survey

The National Anti-Drug Strategy (NADS) Survey series is a series of surveys conducted by the Environics Research Group on behalf of the Government of Canada during the years 2008 – 2012 (Environics Research Group, 2012). An initial baseline survey assessed the behaviours and attitudes of teens aged 13-15 years towards drug use through responses collected from both teens and parents of teens. After this baseline data was collected, Health Canada launched an anti-drug media campaign. Subsequent surveys followed a large proportion of the original respondents as well as new participants to gauge the potential impact of different aspects of the campaign. In this paper we are concerned with deidentification of the baseline survey, which had 1502 respondents.

This dataset is highly sensitive; not due to the population under examination (this is a survey of the general population, and while respondents were adolescents at the time, all participants are now adults), but due to the subject. Survey responses that confirmed acquisition and use of illegal drugs by participants and related to their comfort level in discussing these issues with family members could still impact these individuals and their relationships.

Anonymization of this dataset did not require removal of direct identifiers, as none were provided in the version of the dataset made available to the authors. Potential quasi-identifiers noted within the dataset included subject age, gender, region of residence, household composition (i.e. number of parents, presence of older siblings), aboriginal status and visible minority status. Based on factors including lists of key quasi-identifiers to check for as well as concepts such as attribute persistence (household composition as an adolescent will not persist into adulthood) we decided to focus on age, sex, geographic region, visible minority status and aboriginal status.

Some necessary context: in Canadian government data, aboriginal status is considered an ethno-political category. Survey respondents may be aboriginal or visible minority but are only considered both if they are a mix of aboriginal and some other visible minority group, such as Asian. Visible minority status is coded as a binary variable; respondents who are not Caucasian or aboriginal are considered a visible minority.

In this dataset age had three categories as respondents ranged from 13 to 15, and region had seven categories: one category for Canada's four small Atlantic provinces and six representing each of Canada’s remaining six provinces. Documentation informed us that respondents from the three northern territories were grouped with the geographically nearest province. The remaining variables were binary. Simple multiplication would suggest a total of 168 possible equivalence classes, but since the aboriginal and visible minority statuses are effectively mutually exclusive the actual total is 126 possible equivalence classes. If these were distributed equally across the dataset, we would expect each equivalence class to contain about 12 cases. However, as is usual in practice, they are not distributed equally across the dataset. Some classes are much larger than others. The aboriginal group, for example, made up less than 5 percent of the sample, and the regions are similarly unbalanced with the largest two making up over half the sample. When the equivalence classes were calculated (see appendix for code samples in Stata and R) we found that our dataset had 21 equivalence classes with only a single member, and a total of 42 equivalence classes with less than 5 members. Clearly, even looking at a relatively small number of variables with only a few categories each, in a fairly large dataset, it is very difficult to produce a dataset that satisfies k-anonymity let alone any more stringent criteria. 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.

But how risky would it have been to include the region variable in this survey? This survey is a sample of a much larger population, the population of people aged 13 to 15 in Canada in 2008. Recall that if an equivalence class in the dataset can be assumed to have many co-equivalents in the general population, then membership in an equivalence class cannot be said to reveal confidential attributes or opinions. But how does one know if this is a safe assumption? The first co-author decided to investigate further.

The Census of Canada 2016[3] (Statistics Canada, 2019) is a complete sample of the population of Canada; as such it includes the complete population of people who were age 13 to 15 in 2009, minus any intervening deaths or emigrations. A public use sample is made available for academic researchers to download; although this is a subset of the full data, this sample includes a weight variable that allows it to be weighted back to represent the full population. I decided to use this to determine the size of the general population equivalence classes for our sample unique cases.

It was straightforward to create a version of the Census of Canada dataset that matched the variables and population of the NADS. The Census of Canada also includes questions on gender, visible minority status, and aboriginal identity. The latter two variables needed to be recoded to binary variables, but this could be accomplished unambiguously. The Province / Territory variable was similarly recoded to match the NADS region variable, with the Atlantic provinces being grouped and the populations of Yukon, Nunavut and the Northwest Territories each being assigned to the nearest province. The one variable which posed a slight difficulty was the age variable. NADS respondents would have been age 20 to 22 in 2016; the census microdata file divided age into groups including a category for 20 - 24. I created an artificial variable for ages 20 to 22 by randomly assigning one fifth of the 20 to 24 year old respondents to each of the categories of 20, 21 and 22. The remaining two fifths of the 20 to 24 year old respondents, representing the group that was 23 and 24 years old, were dropped from the dataset along with other respondents outside the targeted age range. This left a nationally representative sample that matched the population from which the NADS was drawn that could be used to estimate the population risk of the variables under consideration.

When I calculated the equivalence classes using the weighted Canadian Census of 2016, 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 our survey was estimated to have a minimum of 369 data twins in the general population. Even though we did not come close to achieving k-anonymity with this set of variables, the sampling factor means that even with the region variable this dataset is low risk, with a population reidentification risk of at most 1/370 for the riskiest case, rather than the sample estimated risk of 1.

The approach outlined here is straightforward to implement and may provide a good option for estimating population risk in appropriate cases. This approach will only work if the sample is drawn from a known population, there exists a survey or census that can be weighted to that population, and that survey contains the correct set of demographic variables. In sample surveys of some clearly defined subset of the general population, a national census may be an appropriate choice, although other large, nationally representative surveys might be used.

Surveys of smaller and more distinct populations that do not approximate a national sample of the general population or a national sample of a readily defined subset of the national population are inherently of higher risk and cannot have their population risk threshold estimated using a national sample. In cases of relatively small samples of a defined population it is reasonable to enforce k-anonymity with a k of 5, while relying on the sampling factor to deal with the concern of revealing opinions or attributes within equivalence classes. If a survey is a complete sample of some defined population (e.g. every person attending a school) or a large fraction of that population then the dataset is of very high risk and should only be shared with extreme caution regarding presence of quasi-identifiers, if it contains no sensitive information, or if the survey respondents were not promised confidentiality.

Drinking Water Quality Survey

In addition to anonymity testing using the type of statistical disclosure risk assessment exemplified by k-anonymity, a researcher may also attempt to check the sensitivity of a dataset to identity disclosure through *penetration testing*, as the second co-author will demonstrate in our next example.

A series of surveys completed by EKOS Research for Health Canada in collaboration with Indian and Northern Affairs Canada during 2007, 2009, and 2011 assessed water quality on First Nations Reserves and rural communities (EKOS Research Associates, 2011). Here we will consider data from the 2007 survey. Demographic data collected from survey participants included their year of birth/age, a binary gender, linguistic preference (English/French), aboriginal status, whether they lived on or off reserve for at least six months of the year, the number of persons in their household, number and age by category of dependent children, number of seniors and vulnerable adults in their household, and whether their home was used as a daycare facility. EKOS also collected information on their place of residence, including whether a drinking or boil water advisory was currently in effect for their region, how many times their community had been under a drinking or boil water advisory during the past five years, forward sortation area (FSA) (a grouping of postal codes often used for releasing census information in Canada), province of residence, rural/urban status, their distance from the nearest city, and the population of their community.

As with the NADS survey, special attention must be taken to prevent re-identification of survey participants, though for different reasons even beyond the usual promise of confidentiality given to survey respondents. In the NADS dataset, full anonymization was particularly important because of the sensitivity of the topic (illegal drugs); here, anonymization is crucial due to the population being considered, as First Nations individuals and communities have been systematically marginalized within Canada. As may be anticipated, these data in combination can be used as key variables by a data intruder to reidentify a survey participant. As the location of First Nations reserves in Canada are publicly known and there is often only a single reserve within a given FSA, a data intruder can easily identify the community of origin for participants who identified as living on-reserve. In the cases for which there are multiple reserves within a given FSA, the data intruder may still discriminate between candidates for the community of origin by comparing the reported population of the community against the 2006 Census records, or juxtaposing reported water advisories against those listed in old news documents, data gathered by civic action groups like Water Today or the Government of Canada's list of long-term water advisories (2020). Even if the data intruder is unable to choose between multiple candidates for a community of origin at this phase though, the populations of First Nations reserves are sufficiently small that individuals demographically unique to the sample will likely be unique to the population of a FSA, especially as the population of individuals living on reserve in a FSA may only number in the tens or hundreds. This data in its original form then threatens the anonymity of survey participants as individuals.

It is also worth noting that even if an individual is NOT sample-unique or population-unique, if all demographically similar individuals can be located to the same reserve, or if all the candidate reserves for these individuals fall under the same tribal governance, locating the participants' community of origin may still pose a problem. Historically, data on First Nations communities has influenced perceptions of them as a group, which is one reason why the research principles of OCAP[4], standing for Ownership, Control, Access and Possession were developed (First Nations Information Governance Centre, 2014); OCAP understands ownership of data as the collective responsibility of the First Nations people from whom it originates. Anonymizing this survey data for public release should then ensure the community of origin for this data cannot be established. A naive attempt to anonymize this survey may assume removal of the FSA to be sufficient. As we will demonstrate, this strategy underestimates the ability of data intruders to use publicly available information and simple programming skills to recover the FSA from still-included information on distance to the nearest city through data linkage. By way of proof, my co-author and I attempted an experiment. She presented me with a dataset from which the FSAs had been suppressed. I constructed a database containing the name, province, population, forward sortation area, location, and distance to the nearest town with population over 15,000 of every First Nations reserve in Canada using resources available to anyone, regardless of academic or governmental affiliations. Information on name, province, and population of First Nations reserves were scraped from Wikipedia pages using Python and the BeautifulSoup code library (Richardson, 2020). Given this information, I then used the Government of Canada's Geolocation webservices (2018) to find the latitude and longitude of these communities. Geonames.org's findNearbyPlaceNameJSON and findNearbyPostalCode functions (Geonames Team, 2020) were used to discover the postal code of each First Nations Reserve, and their straight-line distance to the nearest place with population equal to or greater than 15000 (this being the most appropriate filter accessible through the app).

Having created this database, I then wrote a programming script that would compare the distance-to-the-nearest-city reported by each respondent who had identified as living on reserve, to the distance-to-the-nearest-city for each reserve in that respondent's province within my database. As the distance-to-the-nearest-city recorded in my database was a straight-line distance, while that reported by survey participants was an estimation, an exact match could not be expected. I decided then to create a list of candidate reserves for each survey participant based on whether their estimated distance-to-the-nearest-city came within a given margin of error to the distance-to-the-nearest-city recorded in my database. For 98 of the participants living on-reserve, only a single reserve appeared possible for their location. When my co-author compared my 'guesses' for the FSA of these individuals to the information present in the non-deidentified dataset, 25% of them were correct.

While the correct guesses only amounted to 24 individuals out of 98 supposed correct guesses, out of 1114 individuals surveyed, this experiment demonstrates the risk presented by a data intruder with only simple programming skills and access to public information. Accuracy of the constructed database and its distance-to-the-nearest-city could be improved through use of Geographic Information Systems (GIS) software, which allows measurements and calculations to be made using digital maps. For example, the population size considered to be a city by the Ekos survey was not always consistent with the results returned to me using the geocoding app, which could only limit 'cities' by populations over 5000, 10000, or 15000. An early iteration of the code, using the definition of a city as an area populated with over 5000 people was of low accuracy in predicting participant locations. It is expected that accuracy would improve if a data intruder could select cities of a population closer to that used by the Ekos survey. A wide margin of error had to be used to account for the low accuracy of comparing a straight-line distance-to-the-nearest-city within the constructed database to the estimated distance-to-the-nearest-city by road. GIS layers, such as DMTI route maps, could be used to find the shortest distance by transportation to the nearest city. This would enable us to decrease our allowed margin-of-error, decrease the number of candidate reserves possible for each respondent living on reserve, and increase the number of respondents for which a single location can be positively identified.

Based on the results of this penetration test, the variable 'distance to nearest city' will be dropped from any publicly accessible version of this dataset.

Discussion

The authors' experiences should help illuminate the complexity of determining if a survey dataset has been successfully deidentified and provide some guidance into deidentifying other survey datasets. As a practical approach, steps for deidentification of a dataset might include, first, the removal of all direct identifiers. The set of risky quasi-identifiers to be preferentially retained needs to be identified next. Frequency tables can be used to identify small categories on these quasi-identifiers and determine appropriate groupings. ('Small' is relative and will depend on the size of the sample and the size of the population from which a sample was drawn. As a first pass, groups smaller than 5% of the population might be considered.) Bivariate tables of the grouped quasi-identifiers can be used next to identify variables that produce small groups. (Software such as the program Amnesia or the R package SDCMicro can help automate this process but are beyond the scope of this paper.) The data custodian may wish to consider suppressing individual values rather than regrouping at this stage. For example, in a unique case of a respondent being married and under age 16, the custodian might delete the response to the marriage question instead of regrouping the otherwise non-risky variables of ‘age group’ and ‘marital status’. Larger groups of variables can be iteratively investigated to locate potentially small groupings, until the data custodian comes up with a final set of equivalence classes based on the full list of modified risky variables. If the dataset has achieved k-anonymity with an appropriate value of k (usually 3 or 5) the dataset may be considered provisionally safe. If unacceptably small equivalence classes remain in the dataset and the data custodian would prefer not to drop or regroup variables any further, at this stage the population k-values of the equivalence classes can be checked using an appropriate large national population-weighted dataset, if one can be located. If the dataset has a low population reidentification risk, current variables may be retained, otherwise they will need to be dropped, suppressed or grouped further.

As a final step, variables that relate to geography in any way should be treated with extreme caution. As we have demonstrated, non-obvious geographic variables such as distance from nearest city, combined with contextual information such as survey respondents living on a reservation, can be used to pinpoint geographic location with surprising precision. Other geography-adjacent variables that might need to be considered in relation to contextual survey information might include community size and presence or lack of resources such as a major hospital or public airport in a community. As penetration testing is likely to be beyond what is practical as a part of routine data deidentification, the data custodian should be proactive in considering whether there is a strong analytic interest in retaining such variables and dropping them if there is not.

References

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. Available online: http://hdl.handle.net/10197/9109

Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC) (2018). ‘Draft Tri-Agency Research Data Management Policy For Consultation’. Available at https://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Domingo-Ferrer, J. and Torra, V. (2008). ‘A critique of k-anonymity and some of its enhancements’. Third International Conference on Availability, Reliability and Security, IEEE, pp. 990-993. Available at https://doi.org/10.1109/ARES.2008.97

EKOS Research Associates Inc. (2009). ‘Water Quality On-Reserve Quantitative Research’. Available at http://www.ekospolitics.com/articles/0559.pdf

EKOS Research Associates Inc. (2011). ‘Perceptions of drinking water quality in First Nations communities and general population’. Available at http://www.ekospolitics.com/articles/015-11.pdf

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

Environics Research Group (2012). '2012 national anti-drug strategy (NADS) youth advertising recall and Tracking Survey'. Available at https://epe.lac-bac.gc.ca/100/200/301/pwgsc-tpsgc/por-ef/health/2012/068-11/report.pdf

First Nations Information Governance Centre (2014). 'Ownership, control, access and possession (OCAP™): The path to First Nations information governance'. Available at https://fnigc.ca/sites/default/files/docs/ocap_path_to_fn_information_governance_en_final.pdf

GeoNames Team (2018). 'Geonames'. Available at https://www.geonames.org/export/web-services.html

Government of Canada | Gouvernement du Canada. (2014). Canada's Action Plan on Open Government 2014-16. Available at https://open.canada.ca/en/content/canadas-action-plan-open-government-2014-16

Government of Canada | Gouvernement du Canada. (2018). 'Geolocation service'. Available at https://www.nrcan.gc.ca/earth-sciences/geography/topographic-information/web-services/geolocation-service/17304

Government of Canada | Gouvernment du Canada (2020). 'Ending long-term drinking water advisories', Government of Canada | Gouvernement du Canada. Available at https://www.sac-isc.gc.ca/eng/1506514143353/1533317130660

Richardson, L. (2020). 'Beautiful Soup 4.9.1'. Available at https://www.crummy.com/software/BeautifulSoup

Samarati, P. and Sweeney, L. (1998). 'Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression'. Proceedings of the IEEE Symposium on Research in Security and Privacy, Oakland, CA. Available at https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf

Simi, M S; Sankara, Nayaki and Sudheep Elayidom, M. (2017). 'An extensive study on data anonymization algorithms based on k-anonymity'. IOP Conf. Series: Materials Science and Engineering, 225. Available at https://doi.org/10.1088/1757-899X/225/1/012279

Statistics Canada (2019). '2016 Census of Population [Canada] Public Use Microdata File (PUMF): Individuals File [public use microdata file]', Ottawa, Ontario: Statistics Canada [producer and distributor]. Accessed through ODESI, http://odesi.ca

Thompson, K., (2018). 'Documentation as data rescue: Restoring a collection of Canadian health survey files'. Against the Grain, 29(6), p.12. Available at https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=7876&context=atg

U.K. Economic and Social Research Council (2018). ‘Research Data Policy’. Available at https://esrc.ukri.org/funding/guidance-for-grant-holders/research-data-policy/

U.S. National Science Foundation (2011). ‘Dissemination and Sharing of Research Results’. Available at https://www.nsf.gov/bfa/dias/policy/dmp.jsp

[1]Kristi Thompson is the Research Data Management Librarian at Western University and can be reached by email: kthom67@uwo.ca. Carolyn Sullivan is a student in the Faculty of Information and Media Studies at Western University.

[2] Inferential disclosure - determining from a dataset that having a particular combination of characteristics makes an individual more likely to possess some attribute - is occasionally mentioned in the literature. As forming inferences is the point of most research this is not something that can be eliminated in the general case. More generally the issue of community stigmatization should be considered as a part of the general ethical review of datasets, but this is not precisely a deidentification problem and is beyond the scope of this article.

[3] A long form census of Canada was not conducted in 2011 as it would usually have been, so 2016 was the closest available census occurring after the 2009 survey.

[4] The history of research on First Nations peoples in Canada is complex and deeply problematic. As an official government survey this data on water collection did not fall under OCAP, but we felt the principle of First Nation community rights to privacy should apply when we prepared the public version for deposit.

Appendix: Code for finding equivalence classes in Stata and R

-- Stata --

```
* Stata code for checking k-anonymity
* Kristi Thompson, May 2020

* create the equivalence groups
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
* list the ID numbers of equivalence groups containing 3 or fewer
cases
tab equivalence_group if equivalence_size < 3, sort
* list the values of the quasi-identifiers for each small
equivalence class. E.g. if 1
list var1 var2 var3 var4 var5 if equivalence_group == num
```

--- R --

```
# R code for checking k-anonymity
# Carolyn Sullivan, May 2020

# install plyr, a useful data manipulation package.
install.packages("plyr")
# Load the library.
library('plyr')

datafile <- " location of the data file - csv format -  "
# Read the csv  file.
df <- read.csv (datafile)

# 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)
```