Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

Reconstruction attack risk using Statistics Canada Census Data

by Mathew Abado and George Stefan

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2022
Data Disaggregation: Building a more representative data portrait of society

Reconstruction Attack Risk using Statistics Canada Census Data

Mathew Abado and George Stefan[1]

Abstract

The publication of more disaggregated data can increase transparency and provide important information on underrepresented groups. Developing more readily available access options increases the amount of information available to and produced by researchers. Increasing the breadth and depth of the information released allows for a better representation of the Canadian population, but also puts a greater responsibility on Statistics Canada to do this in a way that preserves confidentiality, and thus it is helpful to develop tools which allow Statistics Canada to quantify the risk from the additional data granularity.

In an effort to evaluate the risk of a database reconstruction attack on Statistics Canada's published Census data, this investigation follows the strategy of the US Census Bureau, who outlined a method to use a Boolean satisfiability (SAT) solver to reconstruct individual attributes of residents of a hypothetical US Census block, based just on a table of summary statistics. The technique is expanded to attempt to reconstruct a small fraction of Statistics Canada's Census microdata. This paper will discuss the findings of the investigation, the challenges involved in mounting a reconstruction attack, and the effect of an existing confidentiality measure in mitigating these attacks. Furthermore, the existing strategy is compared to other potential methods used to protect data – in particular, releasing tabular data perturbed by some random mechanism, such as those suggested by differential privacy.

Key Words: Confidentiality; Reconstruction attack; Differential privacy.

1. Introduction

1.1 Background

Under the Disaggregated Data Action Plan (DDAP), Statistics Canada plans to make a larger amount of disaggregated data available to researchers and the public. The purpose is to ensure that the data better represent the diversity in the Canadian population, including better representation for Indigenous persons, racialized groups, gender diversity, and disability status. Even before the DDAP initiative, Statistics Canada has been publishing a large amount of data at a very disaggregated level. Some of this information is freely available in the data tables that Statistics Canada publishes online. Additionally, Statistics Canada offers several other data access options to researchers, including research data centres, remote data access tools, and data liberation initiatives.

As more information is released and access options are expanded, there is an increased risk of disclosure of individuals' attributes. This disclosure would violate Section 17 of the *Statistics Act*, the Act that governs Statistics Canada and its duties to collect and publish information. One way to examine this risk is by considering a database reconstruction attack. This is an attack where an intruder creates a microdata file consistent with a collection of given statistics, say, those published online, which may then be used to re-identify individual respondents and disclose their attributes. Dinur and Nissim (2003) demonstrated the theoretical risk of such an attack by outlining a procedure using queries to a private database. They found that prevention of this type of attack requires significant perturbation of the data prior to publication.

The effort to address that paper and its implications led to the development of the framework of Differential Privacy. Dwork and Roth (2014) describe a "privacy mechanism" as an algorithm which takes as input a database (e.g., a microdata file) and produces an output (e.g., a cell in a table for publication). This output often corresponds to a query such as "how many rows in the database satisfy property P?", where P might be "resident of Ontario with age between

[1]Mathew Abado, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6 (mathew.abado@statcan.gc.ca); George Stefan, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6 (george.stefan@statcan.gc.ca)

15 and 19", for example. The query result may have disclosure control applied in the form of suppression or perturbation before output. The defining feature of a differentially private mechanism is that it must provide similar outputs from similar input databases. This guarantees confidentiality in the sense that the output will not be able to reveal any single contributor's information, since the output must be similar regardless of the response of any contributor. This will be discussed further in section 3.2.

2. Public Data

2.1 Published Tables

Statistics Canada publishes hundreds of data tables online based on data from each 5-year cycle of the Canadian Census of Population. Additionally, Statistics Canada publishes tables on a wide variety of topics from many sources besides the Census, and indeed the methodology presented in this paper can be extended to analyze these publications as well, but this study will focus exclusively on Census data. These tables are often cross-tabulations of various social characteristics and are available at various levels of detail and levels of geography, down to very fine levels. This amounts to the publication of hundreds of millions of table cells. Statistical disclosure control methods are applied to these tables before publication to protect the confidentiality of the information provided by individual respondents, in accordance with Section 17 of the *Statistics Act*.

The disclosure control methods include, among others, rounding and cell suppression, and will be discussed more in Section 3 of this paper. The aggregation of tables to higher levels of geography or to less detailed variables is also a part of Statistics Canada's disclosure control methodology. If taken too far, however, this approach could limit the potential of the data to represent the experiences of smaller population groups that the DDAP aims to better represent.

With the increasing amount of information disseminated under the DDAP, it is important to continually evaluate the effectiveness of the disclosure control methods applied to these tables. For example, it is important to evaluate the plausibility of an intruder combining the information available in these tables in order to reconstruct the underlying confidential microdata and re-identify individual respondents.

2.2 Expanded Access Options

In addition to the online data tables, several other options are available to researchers who wish to analyze Census data, as well as data from other sources. Research Data Centres (RDCs) are secure environments in which researchers, who have applied for and have been granted clearance, and are thus trusted as deemed employees of Statistics Canada, can access a wide variety of data for their project. Results of their analyses can only be brought outside of the RDC environment for publication after being checked for confidentiality. Real Time Remote Access (RTRA) is an online query tool which is less stringent about access than the RDCs but does not allow the user to directly interact with the microdata. Researchers using the RTRA system can submit queries and receive results in the form of tables with confidentiality rules already applied. The Data Liberation Initiative (DLI) allows post-secondary institutions access to public-use microdata for easier use by students and faculty.

These additional access options increase the amount of information available to a potential intruder, for example, through the results published by researchers based on this data. This increases the risk of disclosure of confidential information. Although each of these access options have independent disclosure control measures in place, these measures are not able to provide any strict mathematical guarantee against disclosure, especially when considering the overall risk of all information being combined. Also, as increasingly more information is released over time, it becomes increasingly difficult to assess the disclosure risks.

2.3 External Data Sources

The Census of Population and its resulting publications are not the only sources of information about the Canadian population. An intruder may have access to additional information about a particular community or person of interest, which could aid them in a reconstruction attack. For example, they may know the genders and ages of their neighbours in a small community and could make educated guesses regarding the range of their salary and other social attributes. An intruder who attempts such an attack may also have access to commercial or administrative databases which could be cross-referenced with their putative reconstructed microdata set for the purposes of re-identification of individuals.

3. Confidentiality Methods

3.1 Rounding and Suppression

Traditional statistical disclosure control methods often involve some form of perturbation. One possible method is conventional rounding, which involves rounding every cell to the nearest base b value (perhaps $b = 5$). For example, if the true number of individuals speaking a particular mother tongue in a Census Subdivision (CSD) is 3, this would be reported as 5. This mechanism, however, is biased in the sense that the expected value of the rounded result is not necessarily equal to the true value; for example, $E[round_5(3)] = E[5] = 5 \neq 3$. An alternative is to perform unbiased random rounding, that is, to round stochastically to either the base value above or below the true value, with probability inversely proportional to the distance between the true value and the rounded value. Using the previous example, a cell value of 3 would be rounded to 5 with probability 0.6 and rounded to 0 with probability 0.4. If the rounding mechanism is known or can be inferred by an intruder, this imposes clear constraints on the underlying microdata which may be used in a reconstruction attack.

Another commonly used disclosure control method is cell suppression. Table cells which are deemed to be too disclosive (e.g., small counts) may be suppressed before publication. Complementary suppression can also be implemented in order to prevent the value of the suppressed cell from being reconstructed using the rest of the table. The intruder is left with a level of uncertainty towards the exact value of the sensitive cell and potentially its complementary cells. If the suppression mechanism is known to the attacker, this can still be a constraint used in the reconstruction process, although it may not be particularly helpful. In this project, suppression has not yet been considered, but the methodology could be expanded to incorporate it as well.

The objective of this paper is to determine the extent to which the rounding mechanisms described above protect data from a potential attacker. In Section 4, several possible attack scenarios will be discussed.

3.2 Differential Privacy

The field of Differential Privacy grew in response to the theoretical risk of a database reconstruction attack demonstrated by Dinur and Nissim (2003) and attempts to provide a more solid mathematical guarantee of privacy. If we consider a mechanism that takes as input confidential microdata and returns a value for publication, then we might consider that mechanism to be private (that is, to protect confidentiality) when its output is not allowed to change substantially as the input data set varies incrementally. The output from such a mechanism would, for example, be similar whether or not any individual contributor was present in the microdata file and is thus unlikely to disclose their contributions. The degree to which the output of the mechanism can vary between adjacent input databases can therefore be used as a measure of its ability to protect the confidential data.

Mathematically, a randomized mechanism M is ε-differentially private if for all possible output sets S in the range of M and for all input databases x, y such that x and y differ in at most one element, we have that $P(M(x) \in S) \leq e^{\varepsilon} P(M(y) \in S)$ (Dwork & Roth, 2014, p. 17). Roughly speaking, any possible output from the mechanism must be approximately as likely whether any given individual is present (say, in database *x*) or absent (say, in database *y*). The value of ε can be thought of as a privacy loss "budget". A larger ε generally allows more data to be released with less perturbation -- the release of data can be thought of as "spending" the privacy budget, and higher quality data costs more.

One example of a differentially private algorithm is the Laplace mechanism, which computes the true results of a set of disjoint counting queries (e.g., a histogram) and perturbs each coordinate with noise drawn from the Laplace distribution with scale parameter $1/\varepsilon$. This mechanism can be shown to preserve ε-differential privacy. As with the rounding strategy described above, the intruder is left with a level of uncertainty as to the precise value of the cell.

4. Database Reconstruction Risk

4.1 Synthetic Data

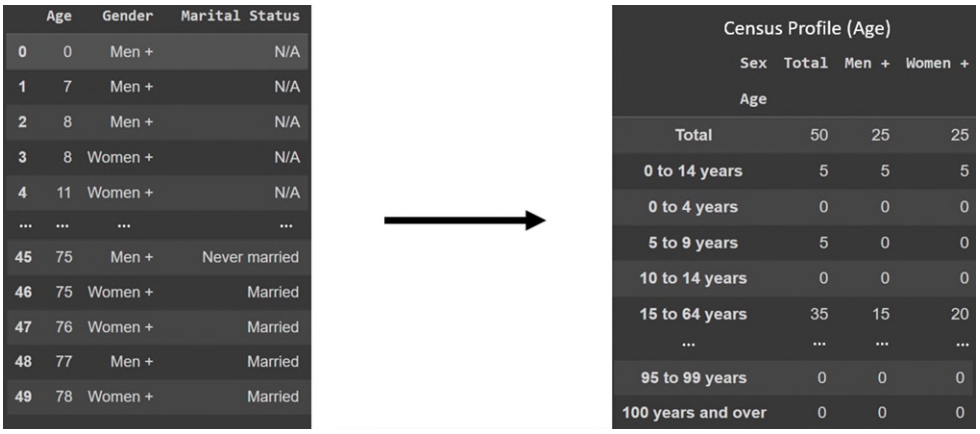
To assess the likelihood of a successful database reconstruction attack on Statistics Canada's published data tables, attack simulations were performed. Synthetic microdata sets were generated representing small CSDs, each with a population of 50. Each simulated person had age (in single years), gender (2-category), and marital status (6-category:

Married, *Living common law*, *Never married*, *Separated*, *Divorced*, *Widowed*) attributes. These attributes were randomly generated according to their joint distribution in the true Canadian population. Attacks were performed on these synthetic datasets instead of on the real data to facilitate repeated tests.

Each microdata set was aggregated to build four contingency tables, in order to mimic tables that are available online at the CSD level for these attributes. The Census Profile is a large, stacked table with population counts broken down by gender alongside many different attributes. For this study, the Census Profile sections concerning age and marital status were recreated; these are effectively one contingency table with gender and age (in 5-year bins) crossed, and another with gender and marital status. Beyond the Census Profile, two more tables were recreated. One of these contained gender crossed with age (in single years) and the other was gender, age (in 5-year bins), and marital status all crossed together. In the case of the latter table, only the six marital statuses listed above were used despite the online table having finer-grained categories.

Unbiased random rounding with base 5 was applied as a confidentiality measure, although the unrounded, i.e., unprotected, versions of these tables were also generated for later comparison.

Figure 4.1-1
Synthetic microdata aggregated and rounded to build the simulated Census Profile (Age section)



	Age	Gender	Marital Status
0	0	Men +	N/A
1	7	Men +	N/A
2	8	Men +	N/A
3	8	Women +	N/A
4	11	Women +	N/A
...	...	...	...
45	75	Men +	Never married
46	75	Women +	Married
47	76	Women +	Married
48	77	Men +	Married
49	78	Women +	Married

Census Profile (Age)

Sex Age	Total	Men +	Women +
Total	50	25	25
0 to 14 years	5	5	5
0 to 4 years	0	0	0
5 to 9 years	5	0	0
10 to 14 years	0	0	0
15 to 64 years	35	15	20
...	...	...	...
95 to 99 years	0	0	0
100 years and over	0	0	0

4.2 Constraint Encoding

A constraint satisfaction problem (CSP) is a problem formulated as a set of variables whose states are related through a set of mathematical and logical constraints, where the goal is to find a variable assignment that is consistent with (i.e., *satisfies)* these constraints. In the case of database reconstruction, each cell in a contingency table represents a constraint on the variables representing personal attributes in the underlying microdata set, and the problem can be formulated as a CSP in this way. Following a procedure outlined by Garfinkel et al. (2019), these constraints were encoded into a CSP format for use in the constraint solver software *Sugar* (Tamura et al., 2009). An example is given in Appendix A.

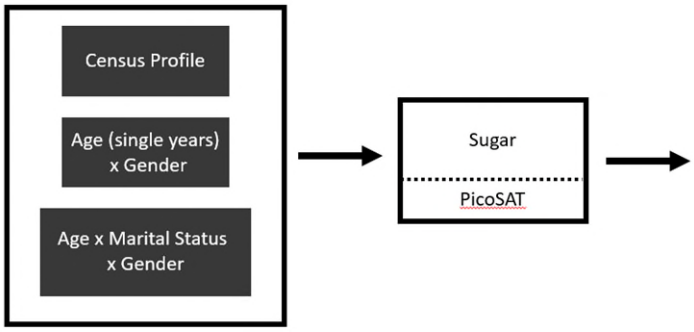
Each underlying person in the microdata set has a variable defined for each of their unknown attributes, in this case age, gender, and marital status. A person contributes 1 to the (unrounded) value of a particular cell if they have the attributes represented by the cell and 0 otherwise, and thus the count can be represented as a sum of these indicator functions. In this way, the value of each cell can be represented by a rather long equation relating, and thus constraining, the individual unknown variables.

Rounding and the resulting uncertainty in both the individual cell counts and in the total population presented additional challenges. The uncertainty in the total population was addressed by including variables for the maximum possible number of underlying people given the reported (rounded) total population. For example, if the reported population was 50, variables for 54 possible people would be defined. An additional binary variable was then included for each potential person to indicate whether or not they truly exist, and these existence variables were also solved for. *Sugar* was then able decide, as part of the solution, how many of these potential people to exclude by “turning off”

their existence variable. The uncertain cell counts were addressed by determining the bounds implied by the rounded values and encoding them as inequalities instead of the equalities (which would have been the constraints implied by known cell counts).

Once these constraints were encoded, *Sugar* could generate a solution by converting the problem into a Boolean satisfiability (SAT) problem and solving that with the help of a SAT solver, in this case *PicoSAT* (Biere, 2008). The resulting solution assigned a value to each of the unknown variables, that is, a microdata set that satisfies all the given constraints and could therefore have plausibly generated the given tables. This is the reconstructed database. An example of the output from *Sugar* is given in Appendix B.

Figure 4.2-1
Schematic representation of the reconstruction of the database corresponding to the microdata from figure 4.1-1. Note the incorrect total population count resulting from the rounded total population in publication.



	Age	Gender	Marital Status
0	2	Women +	N/A
1	4	Men +	N/A
2	5	Women +	N/A
3	9	Men +	N/A
4	12	Men +	N/A
...	...	...	...
50	123	Women +	Widowed
51	124	Women +	Widowed
52	124	Women +	Widowed
53	125	Women +	Widowed

4.3 Match Rate in Reconstructed Databases

Perhaps the most intuitive measure of the success of a database reconstruction, and by extension the risk posed by these sorts of attacks, is the fraction of true microdata records that were successfully replicated in the reconstructed database. This is the *match rate*. It is important to keep in mind that match rates depend strongly on the specific circumstances, such as the population size, number of attributes present, and precise criteria used to confirm a match.

The procedure was performed on 30 synthetic microdata sets, each with 50 people independently generated from the distribution as described in Section 4.1. On average, the corresponding reconstructed databases contained exact matches for 27% (with a standard deviation of $\sigma = 8\%$) of the true population. If the matching criteria are loosened to allow for the reconstructed age to be off by up to 2 years, the match rate increases to 62% ($\sigma = 9\%$).

It is illustrative to compare these match rates to the pairwise match rates between the identically distributed but independent underlying microdata sets. These can be thought of as the incidental match rate between the populations of two different communities, using these criteria for matching. They also approximate what an intruder might be capable of reconstructing given only population statistics — the information that Statistics Canada truly intends to publish — and no additional information. There are 870 pairs between the 30 synthetic microdata sets, and the match rate among these was 15% ($\sigma = 5\%$). With the age match criterion loosened to allow for a 2-year difference, the rate increased to 50% ($\sigma = 7\%$). The fact that these rates are not substantially lower than the reconstructed match rates suggests that many of the reconstructed matches might be accidental.

On the other hand, when the reconstruction procedure was performed instead on the unrounded tables, i.e., without disclosure control measures, the reconstructed databases contained exact matches for approximately 90% of the underlying population. The fact that this is much higher than the match rates obtained using the rounded tables suggests that random rounding is to some degree successful in hindering a reconstruction attack.

4.4 A Very Knowledgeable Intruder

An intruder might, either now or in the future, have access to more information than just that which Statistics Canada publishes online. This additional information might, for example, come from further data disaggregation in the future, from publications stemming from the other data access options we provide (RDC, RTRA, etc.), from knowledge of their local community, or from commercial or administrative databases. This information can be incorporated into a database reconstruction attack. It is hard to know in general how much additional information an intruder might have, but we can model the most extreme case to obtain some risk bounds.

In the most extreme case, an intruder has every value in the microdata set except for one; for example, they might be missing just the age of one individual. It would be instructive to determine whether that last piece of information can be reconstructed using the published tables. This would allow for the risk of a successful full reconstruction to be bounded: if there are multiple possible solutions for the final data point even with all the additional constraints imposed by the additional information, then that ambiguity in the solution must also be present in more reasonable attack scenarios, and the true solution cannot possibly be uniquely determined.

This extreme case is also interesting to consider from the perspective of Differential Privacy. Because the mathematical guarantee provided by the framework specifically guarantees that output will be insensitive to any individual value in the microdata set, an intruder will not be able to reconstruct the last unknown value even in this extreme case. Whatever output is seen by the intruder can plausibly have been generated by several different possible input values, and the true input value can not be reliably chosen from among these.

The additional information corresponding to knowledge of every other microdata cell can easily be encoded as additional constraints in the CSP format for *Sugar*. Each known value is encoded as an additional constraint, and these are appended to the end of the constraints generated from the tables.

Even in this case, reconstruction of the final data point is not guaranteed. If the single missing variable is an age, it can be precisely determined about half of the time. If it is gender, it can be determined in most cases (~75%), which is higher than the baseline rate of ~50% obtained by guessing. If it is marital status, it can be determined in a minority of cases (~25%), which is less successful than a strategy of always guessing that the person in question is married. For comparison, if the tables were not rounded, this extreme attack scenario would always be successful.

5. Discussion

The purpose of this investigation was to evaluate the effectiveness of traditional disclosure control methods, particularly rounding, in the context of initiatives like Statistics Canada's Disaggregated Data Action Plan, which aims to make even more disaggregated data available to researchers and the public. The effectiveness of these methods in protecting the data was evaluated through the lens of database reconstruction attacks, where an intruder uses published data, perhaps with the aid of some additional sources of information and attempts to reconstruct the confidential underlying microdata.

An intruder using data tables without rounding applied for their reconstruction attack was shown to have a high success rate, as measured by the fraction of true microdata records replicated in their reconstructed database (the *match rate*). Once rounding was applied, this match rate was substantially reduced, nearly as low as that which would be expected using only population statistics (i.e., the parameters of the process generating the sample populations, which are the intended objects of study in these tables). Even in the extreme attack scenario where an intruder is imagined to have access to every cell in the microdata table except for one, the rounding performed on the published tables prevented the successful reconstruction of that final data point in many cases. These results are encouraging, but there are some subtleties in their interpretation.

First there is the issue of risk measurement: does the database reconstruction risk framework presented here address every type of disclosure risk, and how should the match rate be interpreted as a measure of that risk? The database reconstruction framework is very flexible, and many kinds of attack can be formulated as constraint satisfaction problems in a similar manner, but not all kinds. It is not clear, for example, how spontaneous recognition could be modeled in this way.

Match rate is an intuitive measure of risk in database reconstruction, but it can be very sensitive to the parameters of the attack. Reconstructions of large populations with fewer, less detailed attributes are likely to have a large number of incidental matches, for example, so it is important to have a baseline for comparison. Interpretation of the match

rate is subtle as well: at what point is it unacceptably high? An intruder who knows that their attack can reconstruct 100% of the true population is a problem, but if they are only able to reconstruct, say, 25% of the population, they might still have no idea *which* of their reconstructed records are correct. This may still be considered a disclosure risk, but additional measures of this risk may be necessary to provide a full picture. Still, even with these limitations and subtleties, match rate could play an important part in a broader disclosure risk assessment protocol.

Furthermore, the simulations presented in this paper were not full-scale database reconstructions of the Canadian population. The scope was limited in population and geographical size, in the number of attributes used, and in the tables used for constraints. An intruder with sufficient expertise, motivation, and access to computing power might be capable of mounting a larger scale attack, and it is possible that the larger number of constraints derived from the additional tables, variables, and nested geographies could allow for higher match rates. This possibility will be especially important to keep in mind if more tables with more variables are made available at more levels of geography, through initiatives like DDAP for example.

Other disclosure control mechanisms are also worth investigating. Differentially private mechanisms are particularly well suited to preventing database reconstruction attacks. For example, the Laplace mechanism can be applied to the published tables and the extent to which this mechanism protects the data from database reconstruction can be determined to see how this compares to the rounding mechanism. The algorithm that was used in the reconstruction attacks in this paper relies on finite bounds for the additive perturbation to each cell, but the Laplace probability density function is unbounded. An intruder might choose artificial bounds, however choosing these bounds to be wide enough to capture the noise but small enough to be informative is fraught. Investigating the match rate of an attempted reconstruction under these conditions would be useful. However, the mathematical guarantee provided by the differentially private mechanism prevents an intruder from being completely sure of any reconstruction: if one reconstructed database is a plausible solution, then any similar database must also be plausible, and thus any particular row (i.e., person) may or may not be a true match. The level of protection offered by other disclosure control mechanisms, differentially private or otherwise, must then be weighed against quality concerns, for example, the ability of the data to represent the needs and diversity of the Canadian population.

References

Biere, A. (2008), “PicoSAT essentials", *Journal on Satisfiability, Boolean Modeling and Computation*, *4*(2-4), pp. 75-97.

Dinur, I., & Nissim, K. (2003, June), “Revealing information while preserving privacy” In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 202-210.

Dwork, C., & Roth, A. (2014), “The algorithmic foundations of differential privacy”, *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), pp. 211-407.

Garfinkel, S., Abowd, J. M., & Martindale, C. (2019), “Understanding database reconstruction attacks on public data”, *Communications of the ACM*, *62*(3), pp. 46-53.

Statistics Act, RSC 1985, c S-49, s 17, https://laws-lois.justice.gc.ca/eng/acts/s-19/

Statistics Canada (2017), *London, CY [Census subdivision], Ontario and Canada [Country]* (table). *Census Profile*. 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. Released November 29, 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E (accessed November 22, 2022).

Statistics Canada (2017), *London, CY [Census subdivision], Ontario and Canada [Country]* (table). *Age (in Single Years) and Average Age (127) and Sex (3) for the Population of Canada, Provinces and Territories, Census Divisions and Census Subdivisions*. 2016 Census. Statistics Canada Catalogue no. 98-400-X2016004. Ottawa. Released May 3, 2017. https://www150.statcan.gc.ca/n1/en/catalogue/98-400-X2016004 (accessed November 22, 2022).

Statistics Canada (2017), *London, CY [Census subdivision], Ontario and Canada [Country]* (table). *Marital Status (13), Age (16) and Sex (3) for the Population 15 Years and Over of Canada, Provinces and Territories, Census Divisions, Census Subdivisions and Dissemination Areas.* 2016 Census. Statistics Canada Catalogue no. 98-400-X2016034. Ottawa. Released August 2, 2017. https://www150.statcan.gc.ca/n1/en/catalogue/98-400-X2016034 (accessed November 22, 2022).

Tamura, N., Taga, A., Kitagawa, S., & Banbara, M. (2009), "Compiling finite linear CSP into SAT", *Constraints*, *14*(2), pp. 254-272.

Acknowledgements

The authors would like to thank our colleagues for their invaluable contributions to the success of this project. In particular, we thank Steven Thomas for his support and guidance at every stage of the project, and for his review of the slides and manuscript. We thank Mark Stinner for sharing his expertise in illuminating discussions, helping us to ask the right questions. We also thank Héloïse Gauvin for her help in reviewing and translating the slides. Finally, we thank Tyler Kirkland and Peter Wright for their review of and suggested improvements to this manuscript.

Appendix A: Sample *Sugar* Code

```
; Excerpt Sugar code demonstrating variable definition and constraint
; encoding.
; Given a reported total population of 50, define variables for 54 people.
; ...
; Define variables to be solved for potential person #53.
(int AGE053 0 125)      ; Unknown age, integer between 0 and 125.
(int GENDER053 0 1)     ; Unknown 2-category gender, encoded as 0 or 1.
(int MARST053 0 5)      ; Unknown 6-category marital status, encoded 0-5.
(int EXISTS053 0 1)     ; Unknown whether person #53 exists, or if there are
                        ; fewer people. Encode 1 for existence, 0 otherwise.
; Example encoding of a cell reporting 5 people between ages 30 and 34.
; Assume that unbiased random rounding is applied, so 5 represents 1-9
; people.

; Insist that the cell count calculated in the inner block is >= 1.
( >=
  ; Add up the indicator variables to obtain cell count.
  ( +
    ; If person #0 exists and is in the age range, add 1, otherwise add 0.
    (if (and (= EXISTS000 1) (>= AGE000 30) (<= AGE000 34)) 1 0)
    ; ... (similar indicator variables for each person)
    (if (and (= EXISTS053 1) (>= AGE053 30) (<= AGE053 34)) 1 0)
  )
  1 ; Insist cell count >= 1.
)
; Insist that the cell count is <= 9 as well.
( <=
  ( +
    (if (and (= EXISTS000 1) (>= AGE000 30) (<= AGE000 34)) 1 0)
    ; ... (similar indicator variables for each person)
    (if (and (= EXISTS053 1) (>= AGE053 30) (<= AGE053 34)) 1 0)
  )
  9 ; Insist cell count <= 9.
```

```
)
```

Appendix B: Sample *Sugar* Output

```
; Excerpt Sugar output with added comments.
s SATISFIABLE     ; At least one solution exists, i.e., an assignment of all
                  ; of the variables consistent with the constraints given.
                  ; Sugar chooses one of these solutions to output.
a AGE000    2     ; Person #0 is reconstructed to be 2 years old.
a GENDER000 0     ; Person #0 is reconstructed with gender encoded as 0.
a MARST000  5     ; Person #0 is reconstructed with marital status 5.
a EXISTS000 1     ; Person #0 is reconstructed to actually exist.
; ...
a EXISTS053 0     ; Person #53 is reconstructed to NOT exist,
                  ; i.e., the reconstructed dataset has fewer than 54 people.
```