Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Statistical Disclosure Control and Developments in Formal Privacy: In Memoriam to Chris Skinner

by Natalie Shlomo

Release date: October 22, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Statistical Disclosure Control and Developments in Formal Privacy: In Memoriam to Chris Skinner

Natalie Shlomo[1]

Abstract

I provide an overview of the evolution of Statistical Disclosure Control (SDC) research over the last decades and how it has evolved to handle the data revolution with more formal definitions of privacy. I emphasize the many contributions by Chris Skinner in the research areas of SDC. I will review his seminal research, starting in the 1990's with his work on the release of UK Census sample microdata. This led to a wide-range of research on measuring the risk of re-identification in survey microdata through probabilistic models. I also focus on other aspects of Chris' research in SDC. Chris was the recipient of the 2019 Waksberg Award and sadly never got a chance to present his Waksberg Lecture at the Statistics Canada International Methodology Symposium. This paper follows the outline that Chris had prepared in preparation for that lecture, and provided to me by his son, Tom Skinner.

Keywords: Risk of Re-identification, Data Revolution, Privacy Models, Differential Privacy

1. Introduction

A special memorial session was held in honour of Chris Skinner at the 2021 Statistics Canada International Methodology Symposium with many moving contributions from friends and colleagues to celebrate Chris' life and achievements. Chris was the 2019 Waksberg Award recipient and was planning on attending the 2019 International Methodology Symposium to deliver his lecture. Unfortunately his illness took a turn for the worse and he sadly passed away on February 21st, 2020. I had the great privilege of presenting this work on Statistical Disclosure Control (SDC), from its early inception and where we are at today with an emphasis on Chris' contributions to the field. The outline of the talk was based on a set of notes that Chris had drawn up in preparation for his 2019 Waksberg Lecture provided to me by his son, Tom Skinner.

In this proceedings paper, I summarize the lecture that I gave in the memorial session. I discuss early SDC developments in Section 2, and move to the situation today based on the Data Revolution in Section 3. Section 4 describes Chris' contributions and his seminal research in SDC. Section 5 presents current research in SDC and data privacy, and one of Chris' final contributions of embedding Differential Privacy into the SDC tool-kit at government agencies. We close in Section 6 with some final words on the impact of Chris' research in government and social statistics and survey methodology.

[1]Natalie Shlomo, Social Statistics Department, University of Manchester, UK, natalie.shlomo@manchester.ac.uk

2. **Early SDC Developments and History**

Since the 1960's there has been public awareness around confidentiality and privacy which initiated public opposition to data collection, particularly for censuses within Europe. For example, there were many objections against the collection of information about the population living in the Netherlands and their last traditional census was held in 1971. This opposition led to a need by government agencies to respond to public concerns about privacy and confidentiality (Dunn 1967) as well as discussed in other early work in Barabba (1975), Cox (1976), Fellegi (1972) and Dalenius (1974). Fellegi (1972, p. 8) wrote: 'National Statistical Institutes (NSIs) live by the good will and trust of the public so that to maintain this trust is literally a question of life or death to them". Based on the research carried out in Sweden, Dalenius (1977) was one of the first to formally define and formalize a framework for Statistical Disclosure Control (SDC) as follows: "An unauthorized party should not be able to learn something about an individual through the release of a statistics f(D) that cannot be learned without access to f(D)".

The work by Dalenius and others provided the framework for researching and developing SDC within government agencies and the establishment of formal governance boards on the release of statistical data. The research being carried out in the United States included, for example the Subcommittee on Disclosure-Avoidance Techniques that was established in 1976 by the Federal Committee on Statistical Methodology, and sponsored by the Statistical Policy Division of OMB as reported in Jabine, et al. (1977) (see also the 1978 report and Appendix A on Statistical Disclosure Avoidance Practices in Selected Federal Agencies). In this Appendix there are five sections with recommendations: the concept of SDC; what to release; disclosure avoidance techniques; effects of disclosure on data subjects and users; and needs for research and development. There were also general rules that were put in place, for example no regional areas that could be published with less than 100,000 individuals.

Further work into the 1980's placed an emphasis on SDC for survey outputs as it was originally and erroneously thought that sampling provided protection against disclosure risks (Dalenius 1988). Paass (1988) was one of the first to estimate the fraction of identifiable records in survey microdata and took into account the sampling and additive noise as well as prior knowledge under an assumed 'attack' on the data. In his paper, Paass (1988) wrote: "Where there is large knowledge, the requirement for privacy protection and high-quality data perhaps may be fulfilled only if the linkage of such files with extensive additional knowledge is prevented by appropriate organizational and legal restrictions." In addition, Bethlehem et al. (1990) was one of the first papers to use probabilistic modelling to estimate the risk of re-identification in survey microdata by estimating the number of population uniques given sample uniques on a set of cross-classified quasi-identifiers. More on this methodology and the contributions of Chris will be presented in Section 4.1.

Into the 1990's there was much more demand for detailed outputs particularly with the availability of better technological solutions and personal computers. There were also rising concerns by users of the data on having to work with protected or perturbed outputs. This coincided with large-scale SDC developments through a scientific evolution of the methodology and the international interchange of theoretical and practical developments, for example, the International Symposium on Statistical Disclosure Avoidance held in the Netherlands in 1990 (as reported in a special issue of Statistica Neerlandica, 1993). There were also cross-collaborations within the European Union through the 4th Framework research project Statistical Disclosure Control(SDC) (1996-1998) and many other EU projects following on from the initial project, including the development of SDC software: mu-ARGUS for microdata and tau-ARGUS for tabular data (specifically cell suppression for magnitude tables containing business statistics). See https://research.cbs.nl/casc/index.htm for more details of the research projects across Europe. A special issue of the Journal of Official Statistics, (Vol 14(4), 1998) titled 'Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data' was particularly impactful and highlighted the large-scale research undertaken in

SDC. In addition, a book and training course were developed (see: Willenborg and De Waal (1996) with contributions by Chris Skinner, and later a second edition in 2001). Continuing work was happening at the same time in the US and Canada, for example the Federal Committee on Statistical Methodology (1994); the Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure (1997); Disclosure Control Issues at Statistics Canada (Yeo and Robertson 1995) including the software package CONFID that also carried out cell suppression for magnitude tables.

Throughout the 1990's, there was growing focus on the development of access and governance arrangements and legislation, and the notion of tiered data access. Data Archives and Research Data Centres were set up along with data governance approaches and frameworks for making effective use of statistical data , for example the '5 Safes' Framework shown in Table 1 and put in practice at the ONS in 2002 (Ritchie 2009) and later the Anonymization Decision-Making Framework (Elliot, et al. 2016 and available at https://ukanon.net/framework/).

Table 1
The 5 Safes Framework

Safe Projects	Is this use of the data appropriate?
Safe People	Can the users be trusted to use the data in an appropriate manner?
Safe Settings	Does the access facility limit unauthorised use?
Safe Data	Is there a disclosure risk in the data itself?
Safe Outputs	Are the statistical results non-disclosive?

3. Data Revolution

Since the latter half of the 2000's, there has been more open and accessible data in the public domain, including open data and big data, leading to greater risks of breaches of privacy and confidentiality since these data sources can potentially be used to compromise released statistical data. In addition, more advanced technological tools are available that enable better data linkages and data manipulation to increase the likelihood of re-identification in statistical data. Government agencies started to become more aware that SDC methods may not be sufficient to protect the confidentiality of statistical units and therefore initiated tighter restrictions and more controlled access to the data. This also manifested in changes to the legislation, particularly, the 2016 EU General Data Protection Regulation (GDPR) which provided provisions and requirements related to the processing of personal data of individuals. There was also more focus on privacy concerns in health data (El Emam, et al. 2011) and genetic data (Homer, at al. 2008, Gymrek, et al. 2013) where the latter were shown to be of high-risk and had implications on the dissemination of DNA databases. In the commercial domain, there were many examples of breaches of privacy: AOL search keywords (Barbaro, et al. 2006), New York City (NYC) taxi trips (Douriez, et al. 2016), Cambridge Analytica and Facebook (Meredith 2018), and others.

With greater technological advancements and the possibility to link data sources, this led to the development of trusted third parties to carry out linkages and secure multi-party computing that was originally developed in the computer science literature and had a cross-over to the statistical literature on how to run advanced statistical modelling under this approach (Slavkovic and Nardi 2007, Snoke et al. 2018). In addition, collaborations between computer scientists and the statistical community grew and led to important developments on database privacy within government agencies (see Section 5 for more details). In the privacy literature, Dwork, et al. (2017) wrote: "Beginning in the mid-2000s, the field of privacy-preserving statistical analysis of data has witnessed an influx of ideas developed some two decades earlier in the cryptography community".

4. Contributions of Chris Skinner to SDC research

Chris's formal research in Statistical Disclosure Control (SDC) started with his collaborations at the University of Manchester to argue for the release of sample microdata (the SARs) from UK Census (Marsh, et al. 1991, Skinner, et al. 1994, Marsh, et al. 1994). This led to his interest on measuring the risk of re-identification in survey microdata through probabilistic modelling first published in Skinner (1992) and described in Section 4.1. He also started his long career of advising for government statistics and data access committees, for example: UK Census Design and Methodology Advisory Committee Statistical Disclosure Control (SDC) Subgroup (2008-2010); Understanding Society Data Access Committee (2010-2013); Expert Advisory Group on Data Access, Wellcome Trust, MRS, ESRC and Cancer Research UK (2012-2014).

4.1 Measuring the Risk of Re-identification in Survey Microdata and Extensions

The disclosure risk scenario for the release of sample microdata containing records from a survey where the sample is drawn randomly from a finite population is based on the following assumptions: (1) there is an 'intruder' (someone with malicious intent to discredit the statistical office) who has access to the microdata and other auxiliary information about the population that allows him/her to link data sources in order to identify individuals in the sample microdata; (2) there is no 'response knowledge' meaning that the intruder does not know who was drawn into the sample of the survey. The basic definition of the risk of re-identification is therefore the probability of correctly being able to make this match. Chris was among the first to develop a statistical modelling framework to estimate the probability of re-identification, conditional on the released data and assumptions about how the data is generated (knowledge of the sampling process). The model is with respect to key variables defined as a set of quasi-identifiers in both data sources and typically categorical such as age, sex, location, ethnic group. Cross-classifying the key variables leads to large contingency tables of sample counts, where many of the cells of the table have a value of zero or a value of one, and we particularly focus on the disclosure risk from the cells of size one, i.e. the sample uniques. The risk of re-identification is based on the notion of population uniqueness in the contingency table: given an observed sample unique in a cell of a table generated from cross-classifying the key variables, what is the probability that the cell is also a population unique? Individual per-record risk measures in the form of a probability of re-identification are estimated. These per-record risk measures are then aggregated to obtain global risk measures for the entire file which are useful to make informed decisions about the level of access.

The probabilistic modelling developed by Chris takes a simplified approach that restricts the information that would be known to intruders (Skinner and Holmes 1998, Elamir and Skinner 2006). Denoting F_k the population size in cell k of a table spanned by key variables having K cells and f_k the sample size and $\sum_k F_k = N$ and $\sum_k f_k = n$. The set of sample uniques, is defined: $SU = \{k: f_k = 1\}$ since these are the potential high-risk records with the potential to be population uniques. Two global disclosure risk measures (where I is the indicator function) are the following:

1. Number of sample uniques that are population uniques:
   $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$
2. Expected number of correct matches for sample uniques assuming a random assignment within cell k (the match probability) $\tau_2 = \sum_k I(f_k = 1)\ 1/\ F_k$

We assume that the population frequencies F_k are unknown and need to be estimated from a probabilistic model where the risk measures are then:

$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(\ F_k = 1|f_k = 1)$ and $\hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}(1/\ F_k\ |f_k = 1)$ (1)

Chris assumed a Poisson distribution and a log-linear model to estimate disclosure risk measures in (1). In this model, he and his co-authors assume that $F_k \sim Pois(\lambda_k)$ for each cell *k*. A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell *k*: $f_k|F_k \sim Bin(F_k, \pi_k)$. It follows that:

$f_k \sim Pois(\pi_k\lambda_k)$ and $F_k|f_k \sim Pois(\lambda_k(1-\pi_k))$ (2)

where the population cell counts F_k are assumed independent given the sample cell counts f_k.

The parameters λ_k are estimated using log-linear modeling. The sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k\lambda_k$. A log-linear model for the μ_k is expressed as: $\log(\mu_k) = x_k{}'\boldsymbol{\beta}$ where x_k is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood (MLE) estimator $\widehat{\boldsymbol{\beta}}$ are obtained by solving the score equations:

$\sum_k(\ f_k - \pi_k \exp(x_k'\boldsymbol{\beta}))x_k = 0$ (3)

The fitted values are then calculated by: $\hat{\mu}_k = \exp(x_k'\widehat{\boldsymbol{\beta}})$ and $\hat{\lambda}_k = \hat{\mu}_k/\pi_k$. Individual disclosure risk measures for cell *k* are:

$P(\ F_k = 1|f_k = 1) = \exp\ (\lambda_k(1-\pi_k))$

$E(1/\ F_k\ |f_k = 1) = (1 - \exp\ (\lambda_k(1-\pi_k)))/\ (\lambda_k(1-\pi_k))$ (4)

Plugging $\hat{\lambda}_k$ for λ_k in (4) leads to the estimates $\hat{P}(\ F_k = 1|f_k = 1)$ and $\hat{E}(1/\ F_k\ |f_k = 1)$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1).

Skinner and Shlomo (2008) develop a method for selecting the main effects and interactions for the log-linear model based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. Defining $h(\lambda_k) = P(\ F_k = 1|f_k = 1)$ for τ_1 and $h(\lambda_k) = E(1/\ F_k\ |f_k = 1)$ for τ_2, they consider the expression:

$B = \sum_k E(I(f_k = 1))(\ h(\hat{\lambda}_k) - h(\lambda_k))$.

A Taylor expansion of *h* leads to the approximation

$B \approx \sum_k \pi_k\ \lambda_k \exp\ (-\lambda_k)(\ h'(\lambda_k)(\hat{\lambda}_k - \lambda_k) + h''(\lambda_k)(\hat{\lambda}_k - \lambda_k)^2/2)$

and the relations $E(f_k) = \pi_k\ \lambda_k$ and $E((f_k - \pi_k\hat{\lambda}_k)^2 - f_k) = \pi_k^2\ E(\hat{\lambda}_k - \lambda_k)^2$ under the hypothesis of a Poisson distribution fit lead to a further approximation of B of the form:

$\hat{B} \approx \sum_k\ \hat{\lambda}_k \exp\ (-\pi_k\hat{\lambda}_k)(\ -h'(\lambda_k)(f_k - \pi_k\hat{\lambda}_k) + h''(\lambda_k)((f_k - \pi_k\hat{\lambda}_k)^2 - f_k)/(2\pi_k))$ (5)

For example, for τ_1:

$\hat{B}_1 \approx \sum_k\ \hat{\lambda}_k \exp(-\hat{\lambda}_k)\ (1-\pi_k)\{\ (f_k - \pi_k\hat{\lambda}_k) + (1-\pi_k)[(f_k - \pi_k\hat{\lambda}_k)^2 - f_k]/(2\pi_k)\}$ (6)

The method selects the model using a forward search algorithm which minimizes the standardized bias estimate $\hat{B}_i/\sqrt{\hat{v}_i}$ for $\hat{\tau}_i, i = 1,2$, which is used as the goodness-of-fit criteria where $\hat{v}_i$ are variance estimates of $\hat{B}_i$. The goodness-of-fit criteria $\hat{B}_i/\sqrt{\hat{v}_i}$ have an approximate standard normal distribution under the hypothesis that the expected value of $\hat{B}_i$ is zero.

Skinner and Shlomo (2008) also address the estimation of disclosure risk measures under complex survey designs with stratification, clustering and survey weights. While the method described assumes that all individuals within cell *k* are selected independently using Bernoulli sampling, i.e. $(f_k = 1|F_k) = F_k\pi_k(1-\pi_k)^{F_k-1}$, this may not be the case when sampling clusters (households). In practice, key variables typically include variables such as age, sex and occupation that tend to cut across clusters. Therefore the above assumption holds in practice in most household surveys and does not cause bias in the estimation of the risk measures. Inclusion probabilities may vary across strata, the most common stratification is on geography. Strata indicators should always be included in the key variables to take into account differential inclusion probabilities in the log-linear model. Under complex sampling, the λ_k can be estimated consistently using pseudo-maximum likelihood estimation (Rao and Thomas 2003), where the estimating equation in (3) is modified as:

$$\sum_k(\hat{F}_k - \exp(x_k'\boldsymbol{\beta}))x_k = 0 \qquad (7)$$

and $\hat{F}_k$ is obtained by summing the survey weights in cell *k*: $\hat{F}_k = \sum_{i\in k} w_i$. The resulting estimates λ_k are plugged into expressions in (4) and π_k is replaced by the estimate $\hat{\pi}_k = f_k/\hat{F}_k$. The goodness-of-fit criteria $\hat{B}$ is also adapted to the pseudo-maximum likelihood method. See Skinner and Shlomo (2008) for a simulation and real application demonstrating this approach for both a simple random sample and a survey with a complex design.

The probabilistic modelling presented here and in other related work in the literature assume that there is no measurement error in the way the data is recorded. Besides typical errors in data capture, key variables can also purposely be misclassified as a means of masking the data, for example through record swapping or the post randomization method (PRAM) (Gouweleeuw, et al. 1998). Shlomo and Skinner (2010) adapt the estimation of the risk of re-identification to take into account measurement errors. Denoting the cross-classified key variables in the population and the microdata as *X* and assuming that *X* in the microdata have undergone some misclassification or perturbation error denoted by the value $\tilde{X}$ and determined independently by a misclassification matrix *M*:

$$M_{kj} = P(\tilde{X} = k|X = j) \qquad (8)$$

The record-level disclosure risk measure of a match with a sample unique under measurement error is:

$$\frac{M_{kk}(1-\pi_k M_{kk})}{\sum_j F_j M_{kj}/(1-\pi_k M_{kj})} \leq \frac{1}{F_k} \qquad (9)$$

Under assumptions of small sampling fractions and small misclassification errors, the disclosure risk measure can be approximated by: $M_{kk}/\sum_j F_j M_{kj}$ or $M_{kk}/\tilde{F}_k$ where $\tilde{F}_k$ is the population count with $\tilde{X} = k$. Aggregating the per-record disclosure risk measures, the global risk measure is:

$$\tau_2 = \sum_k I(f_k = 1)\, M_{kk}/\tilde{F}_k \qquad (10)$$

Note that to calculate the measure only the diagonal of the misclassification matrix needs to be known, i.e. the probabilities of not being perturbed. Population counts are generally not known so the estimate in (10) can be obtained by probabilistic modelling on the misclassified sample as shown above:

$$\hat{\tau}_2 = \sum_k I(\tilde{f}_k = 1)\, M_{kk}\, \widehat{E}(1/\tilde{F}_k | \tilde{f}_k) \quad (11)$$

In more recent work with Chris and presented for the first time in Shlomo and Skinner (forthcoming), a new direction is explored to measure the risk of re–identification for non-probability data sources. More specifically, there are registers in the public domain, where the membership of the register is not known and is sensitive. Examples of registers are of persons with a medical condition, such as Cancer or HIV, or registers that include membership to a loyalty card scheme. The approach can also be extended to the case where samples are drawn from the registers and more generally to non-probability samples, such as those arising from web-surveys. Extending the framework above, the microdata from a random sample can still be used to estimate population parameters under the probabilistic modelling framework for estimating the risk of re-identification, however the complication is to also estimate the propensity of membership for the individuals in the register.

More specifically, let U and U_1 denote the population and the register population, respectively, with $U_1 \subset U$. Let R_i be the register indicator variable for individual i with R_i =1 if $i \in U_1$ and R_i =0 otherwise. As mentioned, we suppose that membership R_i is a sensitive variable for which disclosure is undesirable.

We denote the register population frequencies in cell k by F_k^1. The most risky records are for cells with $F_k^1 = 1$ and, analogous to the derivation presented in Skinner and Shlomo (2008), a risk measure is given by

$$\tau_1^* = \sum_k P(F_k = 1 | F_k^1 = 1) I(F_k^1 = 1) \quad (12)$$

There is no way that these measures can be estimated consistently from the register microdata alone. The microdata provide information about the F_k^1 but not about the F_k in U and distribution of X in U_1 may be quite different to that in U so the microdata carries no direct information about the F_k . Therefore, we use the random sample microdata file in which the values of X are recorded for a probability sample s from U. Let f_k denote the frequency in cell k in s . Note that the f_k and F_k^1 are observed, but the F_k are not. If the intruder has access to the sample microdata file, then it may be advantageous to restrict attention to cells with $f_k = 1$, leading to the following risk measure

$$\tau_1 = \sum_k P(F_k = 1 | F_k^1 = 1, f_k = 1) I(F_k^1 = 1, f_k = 1) \quad (13)$$

Following Skinner and Shlomo (2008), suppose that F_k is Poisson distributed, $F_k \sim Pois(\lambda_k)$ where the parameter λ_k obeys the log-linear model:

$$\log(\lambda_k) = x_k' \boldsymbol{\beta} \quad (14)$$

Suppose that within cell k the unknown membership variable R_i takes the value 1 with probability p_k, independently for each of the F_k units, so that $F_k^1 \sim Pois(\phi_k)$ where $\phi_k = \lambda_k p_k$, and the F_k^1 are binomially distributed $F_k^1 | F_k \sim Bin(F_k, p_k)$ conditional on the F_k. Further, we assume that p_k follows the logistic model:

$$logit(p_k) = x_k' \boldsymbol{\xi} \quad (15)$$

As shown in Shlomo and Skinner (forthcoming), the risk measure τ_1 is estimated by:

$$P(F_k = 1|F_k^1 = 1, f_k = 1) = \frac{\exp(-(1-\pi_k)(\lambda_k-\phi_k))}{1+(1-\pi_k)(\lambda_k-\phi_k)}$$

and to evaluate τ_1^* , we use

$$P(F_k = 1|F_k^1 = 1) = P(F_k - F_k^1 = 0) = \exp(-(\lambda_k - \phi_k)) \quad (16)$$

since $F_k - F_k^1 \sim Pois(\lambda_k - \phi_k)$.

Therefore, the estimation of these measures requires both the estimation of $\boldsymbol{\beta}$ from $f_k \sim Pois(\pi_k\lambda_k)$, and in a second step, the estimate $\boldsymbol{\xi}$, fixing λ_k at the value implied by (14). We then use (16) and the fact that $\phi_k = \lambda_k p_k$ to write

$$log\phi_k = log\lambda_k + x_k'\boldsymbol{\xi} - log\,(1 + \exp(x_k'\boldsymbol{\xi})) \quad (17)$$

and estimate $\boldsymbol{\xi}$ from the fact that $F_k^1 \sim Pois(\phi_k)$ using maximum likelihood estimation and treating λ_k as known. Alternative approaches of estimation are also proposed in Shlomo and Skinner (forthcoming).

Another type of design-based estimator for measuring disclosure risk in sample microdata is called the DIS measure and was developed in Skinner and Elliot (2002) and extended in Skinner and Carter (2003) for more complex survey designs. The disclosure risk is based on a different disclosure risk scenario where an intruder draws a unit at random from the population, checks if the unit is in the sample, and if so, estimates the probability that there will be a correct match to the unit in the sample (this is known as a 'fishing scenario'). Notice that this scenario is quite different than the scenario mentioned under the probabilistic modelling where the intruder has access to a unit in the released microdata and attempts to match the unit to the population. The advantage of this 'fishing scenario' is that the measure can be estimated easily without the need for probabilistic modelling. The DIS measure is defined as

$$\theta = \sum_k I(f_k = 1) / \sum_k I(f_k = 1) F_k \quad (18)$$

and estimated by:

$$\hat{\theta} = \pi n_1 / [\pi n_1 + 2(1-\pi) n_2] \quad (19)$$

where n_1 are the uniques and n_2 are the doubles. Skinner and Shlomo (2012) extend this approach to estimate frequencies of frequencies in finite populations beyond sample uniques.

4.2 Separating Disclosure risk and Harm

Chris provided a conceptual framework in Skinner (2012) for separating potential disclosure

risk from harm, thus linking earlier papers by Duncan and Lambert (1986) and Lambert (1993). The framework is based on decision theory where the actors are the agency, the intruder and the user and they are analysed with respect to their actions and loss functions. Chris emphasized the importance of separating out what can be measured by statistical theory (potential disclosure risk) and what aspects of decision- making requires other inputs, such as policy judgements (potential disclosure harm). This work was also motivated by the Disclosure Risk-Data Utility framework in Duncan, et al. (2001) and the Economics of Privacy in Abowd and Schmutte (2009).

As can be seen from these examples, Chris expanded the depth and breadth of SDC research. Other areas of research where Chris had considerable impact was on the associations between measuring disclosure risks in SDC with other related areas of research, such as record linkage (Skinner 2009) and forensic science (Skinner 2007). Chris' more recent work on disclosure risk and privacy will be the topic of the next sections.

5 Disclosure Risk and Privacy

In the computer science privacy literature, there are more formal definitions of privacy via privacy models that protect against a class of attacks, where the models are parameterized by a threshold of disclosure risk determined a priori through a privacy budget. Generally, the privacy literature requires more perturbative techniques and a greater loss of information to meet the thresholds of the privacy model. The class of attacks are typically based on dealing with inferential disclosure which encompasses both identity and attribute disclosure risks, although the privacy model of k-anonymity (Sweeny 2002) aims to avoid linkage attacks similar to the SDC literature described in Section 4. We note that there is always continuing importance of measuring identity disclosure through the risk of re-identification given the legislation and possible linkage attacks although in the privacy literature there is no distinction between identifying and sensitive variables.

It is important to point out, however, that many concepts in the privacy literature are not new to SDC. For example, the reconstruction attacks mentioned in Garfinkel, et al. (2018) in arguing for more stringent SDC protection for the 2021 US Census have the same considerations as complementary cell suppression developed in the 1980's for protecting magnitude tables of business data, including the calculation of lower/upper bounds on the suppressed cells. Indeed, the reconstruction attack is not about linkage rather it is concerned with attribute disclosure through small cell counts, particularly on the margins. As mentioned the privacy literature mainly focuses on attribute and inferential disclosures although the SDC literature have also covered these topics, for example the predictive disclosure risk mentioned in Fuller (1993). Another privacy model is tracing attacks where one can infer whether an individual is in a sensitive dataset, eg. Homer et al. (2008), but the SDC literature has also focused on whether a data subject is visible in the dataset. The difference between the privacy literature and the SDC context is that the statistical unit consents to providing their data for statistical purposes and the government agency thus has a legal and ethical obligation to protect against breaches of disclosures.

Since 2005, there have been four collaborative meetings between the SDC community and the computer science privacy community and this has led to substantial understanding of the different approaches both with respect to guarantying privacy and maintaining sufficient utility in the data. For example, in Nissim, et al. (2017, p. 5), it is mentioned: "Privacy is a property of an informational relationship between input and output not a property of output alone", and this has led to some relaxations of the strict privacy guarantees in the privacy literature. On the other hand, the SDC community have recognized the need to have more formal privacy guarantees, particularly with the demand to allow for accessing the statistical data via web-based dissemination applications. The collaborations between the SDC community and the computer science privacy community have also led to a journal that was initiated in 2005, titled the Journal of Privacy and Confidentiality (https://journalprivacyconfidentiality.org) of which Chris served as one of the first co-editors (Abowd, et al. 2009).

5.1 Differential Privacy

Dwork and Naor (2010) show that the Dalenius (1977) definition of a privacy breach presented in Section 2 is impossible to prevent and proposed that instead of comparing information with and without f(D), they compare f(D) and f(D') where D' is the database D without a single unit. The privacy model is known as Differential Privacy (Dwork et al. 2006).

In Differential Privacy, a 'worst case' scenario is allowed for, in which the potential intruder has complete information about all the units in the database except for one unit of interest. The definition of a perturbation mechanism M satisfies ε-differential privacy if for all queries on neighbouring databases $D, D' \in A$ differing by one individual and for all possible outcomes defined as subsets $S \epsilon Range(M)$ we have:

$$p(M(D)\epsilon S) \leq e^{\varepsilon} p(M(D')\epsilon S)$$

A relaxation is offered by the definition of (ε, δ)-differential privacy:

$$p(M(D)\epsilon S) \leq e^{\varepsilon} p(M(D')\epsilon S) + \delta$$

This means that observing a perturbed output S, little can be learnt (up to a degree of e^{ε}) and the intruder is unable to determine whether the output was generated from database D or D'. In other words, the ratio $p(M(D)\epsilon S)/$ $p(M(D')\epsilon S)$ is bounded and the probability in the denominator cannot be zero. Thus, Differential Privacy formally bounds increased disclosure risk from participating in the database. Under the(ε, δ)-differential we allow a small amount of slippage to this constraint.

The solution to guarantee Differential Privacy in the computer science literature is by adding noise/perturbation to the outputs of the queries under specific parameterizations and in the privacy literature the noise is generated from the Laplace Distribution (and for count data, a discretized Laplace Distribution can be used).

Shlomo and Skinner (2012) first looked at whether standard SDC methods are differentially private mechanisms according to the definition above. They found that sampling as an SDC method is not differentially private since a person may be observed in the (protected) sample but if the person is removed from database D to obtain D', this causes an impossible situation where the sample count is greater than the population count. In fact, any non-perturbative SDC method, such as coarsening variables, is not differentially private because one can always find a case where the denominator in the ratio is zero due to the deterministic nature of the data protection. Perturbative methods in the SDC tool-kit can be made differentially private if the perturbation mechanisms do not have zero probabilities of perturbation. As an example, SDC methods traditionally do not perturb zero cells in census tables containing whole population counts, rather stochastically induces more zeros through the perturbation, for example through random rounding. However, to make this perturbation approach differentially private, the (random) zeros of the table also need to be perturbed.

5.2 Online Flexible Table Builders

There has been much interest by government agencies to develop online flexible table builders for generating census tables which allows users to define and download their own census tables, typically through a dedicated website with a predefined set of variables and their categories selected through drop-down lists. Light disclosure checks are carried out on the generated tables prior to their release. One such application was developed at the Australian Bureau of

Statistics (ABS) (see: https://www.abs.gov.au/statistics/microdata-tablebuilder/tablebuilder). The application uses a perturbation vector to change values of cell counts depending on the original cell value, where the perturbation mechanism have the properties of being bounded, unbiased, have maximal entropy, only allow for non-negative perturbations and zero cells are not perturbed. Shlomo and Young (2008) introduced an approach to transform the perturbation vectors in such a way that the marginal counts are preserved in expectation by introducing the property of invariance into the perturbation mechanism.

In the ABS online flexible table builder, a small random number is assigned to each individual in the census microdata. Then, when a table is requested and the individuals are aggregated into the cells of the table, the random numbers of the individuals in each cell are also aggregated. This aggregated random number is then used as the seed to determine the perturbation (Fraser and Wooton 2005). This means that any time a same cell appears in any requested census table, it will always have the same perturbation. Therefore, there is no risk of being able to 'unpick' a true cell value by averaging out independent perturbations under multiple requests of the same table. In addition, this approach ensures that the perturbation is what is known as a 'non-interactive' mechanism since essentially all outcomes of perturbation on requested census tables within the online flexible table builder are known in advance.

One of Chris' last initiatives prior to his illness was to take the lead on setting up a collaborative programme between statisticians, computer scientists, social scientists and practitioners held at the Isaac Newton Institute, University of Cambridge. Together with Professor David Hand, they successfully launched the Data Linkage and Anonymization Programme (supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/K032208/1) held from July to December 2016. It was during this programme that a group of statisticians looked at whether Differential Privacy could be a viable solution for an online flexible table builder for generating census tables, resulting in the paper by Rinott, O'Keefe, Shlomo and Skinner (2018). The main difference with the original ABS approach was to use a differentially private perturbation mechanism (known as the Exponential Mechanism which is essentially a discretized Laplace Distribution) and to perturb the (random) zero cells. Any resulting negative perturbations were then pushed to zeros in the census tables. Assuming independent pertubations, the Exponential Mechanism is defined as follows: for a given cell count value a, choose $b \epsilon B$ (where B is the range of b) with probability proportional to $\exp\left(\frac{\left(\frac{\varepsilon}{2}\right)u}{\Delta u}\right)$ where u is the perturbation and Δu is the maximum difference of a cell count in database D versus D', which for the case of a census table of internal cells, take the value of one. Accounting for marginals in the census tables raises the complexity of the perturbation vector (see Rinott, et al. 2018 for more information about marginals). In order to ensure utility, the perturbations were capped at ± 7 thus the mechanism satisfied (ε, δ)-differential privacy. An example of a pertubation vector for $\varepsilon = 1.5$ and $\delta = 0.00002$ and a perturbation cap of ± 7 is the following:

u	-7	-6	-5	-4	-3	-2	-1	0
$p(u)$	0.00002	0.00008	0.00035	0.00157	0.00706	0.03162	0.14172	0.63516

u	1	2	3	4	5	6	7
$p(u)$	0.14172	0.03162	0.00706	0.00157	0.00035	0.00008	0.00002

Examples and applications are shown in Rinott, et al. (2018) and in addition, they show how to adjust statistical analyses when carried out on the perturbed data given that the perturbation mechanism is known and not secret under Differential Privacy.

Differential Privacy, with more formal by-design privacy guarantees to protect against attribute and inferential disclosure risks, may provide solutions to protect statistical data when disseminated as open data and via web-based internet applications and may become part of the SDC tool-kit within government agencies. The US Census Bureau will be applying Differential Privacy in their 2021 census products (Abowd 2018). Further research is needed on how privacy budgets are influenced when combined with other SDC approaches, such as coarsening, sampling and variable suppression. There is also ongoing research within the privacy literature to improve the utility of differentially private perturbation mechanisms, for example bounded Differential Privacy in Kifer and Machanavajjhala (2014).

6. Final Words

In summary, a key feature of Chris's approach to research on SDC, as well as his other areas of research, was that it was based on finding practical solutions to real statistical problems. His research was influential because he was able to put theory to practice and to solve real problems to advance the social sciences, government and social statistics and survey methodology. His decades of research in SDC and other research areas in survey statistics, including missing data and measurement error, data integration, the analysis of complex survey designs, multiple frame estimation and more, made him the definitive voice of a generation.

References:

Abowd, J. M. (2018), The U.S. Census Bureau Adopts Differential Privacy. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2867. https://doi.org/10.1145/3219819.3226070

Abowd, J.M. and Schmutte, I.M. (2019), An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. American Economic Review, 109(1), 171-202.

Abowd, J.M., Nissim, K. and Skinner, C.J. (2009), First Issue Editorial. Journal of Privacy and Confidentiality 1 (1).

Barabba, V.P. (1975), The right to privacy and the need to know. In: US Bureau of the Census: A numerator and denominator for measuring change. Technical Paper 37.

Barbaro, M. and Zeller Jr, T. (August 9, 2006), A face is exposed for AOL searcher no. 4417749. New York Times.

Bethlehem, J.,Keller, W. and Pannekoek, J. (1990), Disclsoure Control of Microdata. JASA 85, 38-45.

Cox, L.H. (1976), Statistical disclosure in publication hierarchies. Report No. 14 of the research project Confidentiality in Surveys, Dept. of Statistics, University of Stockholm.

Dalenius, T. (1974), The invasion of privacy problems and statistics production-an overview. Statistisk tidskrift 3, 213-225.

Dalenius, T. (1977), Towards a methodology for statistical disclosure control. Statistik Tidskrift 15, 429-444.

Dalenius, T. (1988), Controlling Invasion of Privacy in Surveys. Statistics Sweden.

Douriez, M., Doraiswamy, H. Freire, J. and Silva, C.T. (2016), Anonymizing NYC Taxi Data: does it matter? IEEE internatinl Conference on Data Science and Advanced Analytics (DSAA2016), 140-148.

Duncan, G. and Lambert, D. (1986), Disclosure-limited data dissemination (with discussion). JASA, 81, 10-28.

Duncan, G., Keller-McNulty, S., and Stokes, S. (2001), Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. Technical Report LA-UR-01-6428. Statistical Sciences Group,Los Alamos, N.M.:Los Alamos National Laboratory.

Dwork and Naor (2010), On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. Journal of Privacy and Confidentiality.

Dwork, C. Smith, A.., Steinke, T. and Ullman J. (2017), Exposed! A survey of attacks on private data. Annual Review of Statistics and its Application. Harvard University.

Dunn, E.S. (1967), The idea of a national data centre and the issue of personal privacy. Am. Statistician, 21, 21-27.

Elamir, E. and Skinner, C.J. (2006), Record-level measures of disclosure risk for survey microdata. Journal of Official Statistics, 22, 525-539.

Fellegi, I.P. (1972), On the question of statistical confidentiality. JASA, 7-18.

Fraser, B. and Wooton, J. (2005), A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing. Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, 9-11 November.

Fuller, W.A. (1993), Masking Procedures for Micro-data Disclosure Limitation. JOS 9, 383-406.

Garfinkel, S., Abowd, J.M., Martindale, C. (2018), Understanding Database Reconstruction Attacks on Public Data. ACM QUEUE 16 (5)

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998), Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, 14, 463-478.

Gymrek, M., McGuire, A.L., Golan, D. et al. (2013), Identifying personal genomes by surname inference. Science, 339 (6117) 321-324.

Homer, N. Szelnger, M., Redman, D. et al. (2008), Resolving individuals contributing trace amoounts of DNA to highly complex mixures using high-density SNP genotyping microarrays. PLoS Genet, 4(8).

Hundepool, A., et al. (1998), mu-ARGUS user's manual and tau-ARGUS User's Manual. Department of Statistical Methods, Statistics Netherlands, The Netherlands.

Jabine, T.B., Michael, J.A., Mugge, R.H. (1977), Federal Agency Practices for Avoiding Stastical Disclosure: Findings and Recommendations. Asasrms.org.

Kifer, D. and Machanavajjhala, A. (2014), Pufferfish: A framework for mathematical privacy definitions. ACM Transactions on Database Systems, 39(1), 1-36.

Marsh, C., Dale, A. and Skinner, C.J. (1994), Safe data versus safe setting: Access to microdata from the British Census. International Statistical Review, 62, 35-53.

Marsh, C., Skinner, C.J. and 6 co-authors (1991), A case for samples of anonymised records from the 1991 Census. Journal of the Royal Statistical Society, A, 154, 305-340.

Meredith, S. (April 10, 2018), Facebook-Cambridge Analytica: A timeline of the data hijacking scandal". CNBC.

Nissim, K., Steinke, T. Wood, A., Altman, M. and 5 other authors (2017), Differential Privacy: A Primer for a Non-technical Audience. Available at: https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_0.pdf

Paass, G. (1988), Disclosure risk and disclosure avoidance for microdata, Journal of Business and Economic Statistics, 6(4), 487-500.

Rao, J. N. K., and Thomas, D. R. (2003), Analysis of categorical response data from complex surveys: An appraisal and update. In Analysis of Survey Data (eds R. L. Chambers and C. J. Skinner), pp. 85–108. Wiley, Chichester, UK.

Rinott, Y., O'Keefe, C., Shlomo, N., and Skinner, C. (2018), Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. Statistical Sciences, Vol. 33, No. 3, 358-385.

Ritchie, F. (2009), Designing a national model for data access. Comparative Analysis of Enterprise (Micro) Data 2009.

Skinner, C.J. (1992), On identification disclosure and prediction disclosure for microdata. Statistica Neerlandica, 46, 21-32.

Shlomo, N. and Skinner, C.J. (2010), Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. Annals of Applied Statistics, 4(3), 1291-1310.

Shlomo, N. and Skinner, C.J. (2012), Privacy Protection from Sampling and Perturbation in Survey Microdata, Journal of Privacy and Confidentiality, Vol. 4, Issue 1.

Shlomo, N. and Skinner, C.J. (forthcoming), Measuring Risk of Re-identification in Microdata: State-of-the Art and New Directions. To be published: Journal of the Royal Statistical Society, Series A.

Shlomo, N. and Young, C. (2008), Invariant Post-tabular Protection of Census Frequency Counts. In PSD'2008 Privacy in Statistical Databases, (Eds. J. Domingo-Ferrer and Y. Saygin), Springer LNCS 5261, 77-89.

Skinner, C.J. (2007), The probability of identification: applying ideas from forensic statistics to disclosure risk assessment. Journal of the Royal Statistical Society, Series A, 170, 195-212.

Skinner, C.J. (2009), Record linkage, correct match probabilities and disclosure risk assessment. In Insights on Data Integration Methodologies: ESSnet-ISAD workshop, Vienna, 29-30 May 2008, Eurostat Methodologies and Working papers, Luxembourg, European Communities, 11-23.

Skinner, C.J. , Marsh, C., Openshaw, S., and Wymer, C. (1994), Disclosure Control for Census Microdata. Journal of Official Statistics 10, 31-51.

Skinner, C.J. and Holmes, D. (1998), Estimating the re-identification risk per record in microdata. Journal of Official Statistics 14, 361-372.

Skinner, C.J., and Elliot, M. J. (2002), A measure of disclosure risk for microdata. Journal of the Royal Statistical Society, B, 64, 855-867.

Skinner, C.J. and Carter, R.G. (2003), Estimation of a measure of disclosure risk for survey microdata under unequal probability sampling. Survey Methodology 29, 177-180.

Skinner, C. J. and Shlomo, N. (2008), Assessing identification risk in survey microdata using Log Linear models. JASA 103 (483), 989-1001.

Skinner, C.J. (2012), Statistical disclosure risk: separating potential and harm. International Statistical Review, 80, 349-368, with discussion and rejoinder 379-391.

Skinner, C.J. and Shlomo, N. (2012), Estimating Frequencies of Frequencies in Finite Populations. Statistics and Probability Letters Vol. 82, 2206-2212.

Slavkovic, A.B., Nardi, Y. and Tibbits, M.M. (2007), Secure logistic regression of horizontally and vertically partitioned distributed databases. Seventh IEEE International Conference on Data Mining Workshops, ICDMW2007), 723-728.

Snoke, J., Brick, T. Slavkovic, and A. Hunter, M.D. (2018), Providing accurate models across private partitioned data: Secure maximum likelihood estimation. Annals of Applied Statistics 12(2), 877-914. Statistica Neerlandica (1993). Special Issue: Proceedings of the International Symposium on Statistical Disclosure Avoidance. Volume 46, No. 1.

Sweeney, L. (2002), k-anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 557-570.

Willenborg, L. and De Waal, T. (1996), Statistical Disclosure Control in Practice, Lecture Notes in Statistics, 111. New York: Springer Verlag.

Willenborg, L. and De Waal, T. (2001), Elements of Statistical Disclosure Control in Practice. Lecture Notes in Statistics, 155. New York: Springer-Verlag.

Yeo, D. and Robertson, D. (1995), Disclosure control issues at Statistics Canada. https://publications.gc.ca/collections/collection_2017/statcan/11-613/CS11-617-96-5-eng.pdf