Disclosure Risk and Data Quality

Cox, Lawrence
National Institute of Statistical Sciences
12177 Etchison Road
Ellicott City, MD 21042 USA
cox@niss.org

Introduction

National Statistical Institutes (NSIs) have the dual responsibility to release high quality data while limiting the risk of disclosure of confidential subject information. In the past, the disclosure limitation was done in a mostly ad hoc manner—namely, some form of disclosure limitation, possibly heuristic or ad hoc, is applied until the NSI is comfortable that disclosure risk is acceptable. Currently, most NSIs have available one or more means of principled disclosure limitation. Still, in most cases quality review is only partial. In other cases, data may be released without any quality review. In the typical case today, disclosure limitation is adequate, but perhaps overly protective, and the quality review amounts to manually comparing statistical outputs from original and disclosure limited (masked) data to assure fidelity of a handful of key estimates. If the estimates are conformal, the masked data are released. If not, the masked data are modified (tweaked) manually until the NSI is comfortable with both disclosure protection and data quality. This approach is highly subjective and difficult to impossible to reproduce or measure or to compare alternative disclosure-quality treatments. To move balancing confidentiality with quality from art to science and to improve reliability, reproducibility, data security and data quality, what is needed are (1) quantitative measures of disclosure risk and data quality, (2) functional forms measuring the balance between these measures, and (3) disclosure limitation methods that are sensitive and responsive to both the risk and quality measures. These are ambitious goals (Cox, Karr and Kinney 2011), and their difficulty is evident when standard rules and procedures are examined from the point of view of measuring disclosure risk (only). We undertake this examination, illustrate some issues raised for the case of tabular data and provide simple examples.

Measuring Disclosure Risk and the Risk-Quality Tradeoff for Tabular Data

We focus on the two dominant types of tabular data: *contingency table* (*count*) *data* and *magnitude data*.

Count data

For count data, the value of a tabulation cell equals the number of subjects in the population or sample with characteristics matching the group characteristics defining the cell, e.g., number of university professors residing in a particular geographic area and age 40-50 years. Disclosure in count data has been quantified in two ways. First, if *small* nonzero counts are presented or can be inferred from released data, the likelihood that an individual can be identified and confidential information revealed is considered to be high. This is referred to as a *t-threshold rule*: a released or inferred tabular cell is a *disclosure cell* (is small) if the cell count *n* satisfies $0 < n < t$, where *t* is a predetermined threshold . Threshold rules are widely used by NSIs. The reasoning supporting this definition is that greater anonymity is afforded members of larger as opposed to smaller groups. If the NSI assumes that intruder knowledge is restricted to the released tabulations and membership of an individual in a population group (of size *n* subjects), then the intruder can associate the individual with a subgroup characteristic of size *s* with *reidentification probability* at most *s/n*. This type of attack is referred to as a *targeted intrusion*. If instead the intruder is on a *fishing expedition*—looking to identify or disclose confidential information for any individual, e.g., to gain attention or embarrass the NSI—then small counts are the logical place to begin an intrusion, viz., to seek additional information--within or beyond the tabulations—to identify and discover confidential information pertaining to an individual subject. A fishing expedition is obverse to a targeted intrusion in the sense that in a fishing expedition the intruder proceeds from the general (a small count) to the specific (an individual and its characteristics) instead of from the specific (an individual) to the general (a characteristic exhibiting a small count).

Typically, zero counts are not considered "small" for confidentiality purposes because zeroes are often known in advance from general knowledge or otherwise easy to identify. This is consistent with the preceding discussion, as zero counts have zero probability of disclosure. Zero counts can be important for analytical purposes, e.g., in public health data, in which case it is desirable to exempt zeroes from the effects of disclosure limitation wherever possible. Preserving zero counts, then, is our first example of a *disclosure risk-quality interface* (or *tradeoff*).

The second kind of disclosure occurs when all counts within a group are clustered within a single category, as this reveals that every member of the group possesses the trait associated with this category, viz., the associated probability equals 1. Similarly, if the category dominates the subgroup counts, this probability is nearly 1. This is known as *group disclosure*.

We assert that for targeted intrusion disclosure risk is measured by the reidentification probability *s/n*, whereas for a fishing expedition it is measured by *n*. Consequently, a *t*-threshold rule will adequately address disclosure risk posed by a fishing expedition, but fails to do so against targeted intrusion or group disclosure. This is illustrated by simple examples, based on $t = 5$. If $n = 100$, $s = 90$, and all other subgroup counts are either 0 or at least 5, then the *t*-threshold rule detects no disclosure, even though the reidentification risk is 0.9. Conversely, if $n = 100$, $s = 2$ and all other nonzero subgroups are large, then, assuming that the intruder has group but not subgroup knowledge, the reidentification risk is 0.02 and not worthy of disclosure limitation against targeted intrusion. If, on the other hand, the intruder has subgroup knowledge, then disclosure was present prior to the tabulations and irrespective of the size *s*. The second paradigm is even simpler: if $s = n \geq t$, then group disclosure occurs with probability 1 but no disclosure is detected.

Another weakness of the *t*-threshold rules was recently revealed through mathematical research. Implicit in the selection of the threshold *t* defining small values is the assumption that all values 1, 2, …,t-1 can be realized in the tabulations because, otherwise, the number of actual possibilities would be too few and reidentification risks unreliable. DeLoera and Onn (2004) showed that there can be *gaps* in the sequence of permissible integer values for table cells, calling into question the appropriateness of the t-threshold rule.

Inadequacy of the *t*-threshold rule is easily understood. The *t*-threshold rule was established decades ago at a time when table preparation and disclosure limitation were performed by hand--before the age of computers. Consequently it is a simple rule that can be employed in a paper and pencil manner. It addresses attack via fishing expedition—which was a real threat at that time—but not attack via targeted intrusion—which was far less likely to be attempted in the absence of computers, matching methodologies, etc. Curiously, during the period before computers and modern statistical disclosure limitation (*SDL*) methods, it would probably have been simpler to identify and address (heuristically) group disclosure, but there is to my knowledge no indication that NSIs attempted to do so. To this day, the disclosure limitation procedures of NSIs tend to ignore group disclosure.

We conclude that more than a *t*-threshold rule is necessary to quantify disclosure risk in count data. Enhancing current definitions of disclosure risk in count data must be done in cognizance of the SDL methods that may be employed—or point to the need for new methods. The major obstacle posed by current SDL methods for count data is that they are intertwined with the *t*-threshold rule in terms of cell size (*s* or *n*) and the additive structure of the tabulations, and thus may not readily accommodate a percentage-based (*s/n*) definition of disclosure risk. New methods must also deal with integer gaps.

Standard post-tabular disclosure limitation methods for count data subject to a *t*-threshold rule include the following. *Primary cell suppression*—suppressing all disclosure (small) cells—followed by *complementary cell suppression*—suppressing additional, nondisclosure cells until values of small cells cannot be deduced or narrowly estimated from the set of unsuppressed tabulation cells. *Rounding* all tabulations (counts) to rounding base *t* so that small cells are not presented and consequently cannot be narrowly inferred. *Randomly perturbing* counts is the third method.

To assure that the rounding cannot be undone (a confidentiality concern) and that rounded tabulations are additive (a quality and usability concern), rounding is performed in a more general manner than the *conventional rounding* procedure we learned as children. Although conventional rounding is *minimum distance rounding*, viz., rounded counts are multiples of the base closest to original counts, it is nonadditive: $3 + 4 = 7$ but, base 5, $5 + 5 \neq 5$. Additive rounding is known as *controlled rounding* (Cox and Ernst 1982); controlled rounding can be deterministic or stochastic (Cox 1987). However, beyond two-dimensional and related tables, controlled rounding is not always assured (Cox and George 1989) and relaxed definitions or heuristic approaches are needed. Similar statements are true for random perturbation.

Partial rounding is also possible. Partial rounding can be viewed as a specific form of *controlled tabular adjustment* wherein small counts are replaced by either 0 or *t* and other counts are adjusted slightly using mathematical programming to restore additivity.

Cell suppression is deleterious to data quality (Cox 2008), and can be vulnerable to intruder attack (Cox 2009a). Rounding, perturbation and controlled tabular adjustment are designed all assure confidentiality and to preserve two facets of data quality—(1) assuring that, subject to confidentiality requirements, released counts remain as close as possible to original counts and (2) preserving additivity of the tabular structure (Cox and Kim 2006; Cox, Orelien and Shah 2006; Cox 2009b). These considerations constitute our second and third disclosure risk-quality interfaces. The first full, quantitative disclosure risk-data quality assessment was done for rounding (Cox and Kim 2006), where it was shown that a zero-restricted (preserve multiples of the base), 50/50 rounding rule (round nonmultiples up or down with probability ½) performed best in terms of data quality and led to a uniform posterior predictive distribution of original counts conditional on (released) rounded counts—perfect from a disclosure risk standpoint.

Magnitude data

For magnitude data, tabulation cells are defined by subgroup characteristics and a statistic of interest. The cell value equals the sum over all subjects with characteristics matching the subgroup characteristics of each subject's value for the statistics of interest, e.g., total value of shipments of all manufacturing establishments located in a particular geographic area, with 100-500 employees, and assigned a particularly industrial activity (NAICS) code. Individual subject values are called *contributions* to the cell value. Count data can be viewed as magnitude data for which each subject contributes 1 to the cell value if its characteristics match the subgroup characteristics. Disclosure in magnitude data typically has been defined by a *dominance rule* such as the *p-percent rule*: if the total contribution of all but the largest and second largest contributors is less than *p*-percent of the largest contribution, the cell is a disclosure cell. This rule protects the largest contributor from narrow estimation (*p*-percent or less) of its value by the second largest—and consequently protects every contributor from every other contributor by at least *p*-percent.

Cox (1981) provides a general theory for a large class of disclosure rules that includes threshold and dominance rules based on *linear sensitivity measures*. This theory enables the NSI to compute a lower bound on the amount of disclosure protection needed to protect each individual disclosure cell. As rounding and perturbation are generally ineffective for the large values and skewed distributions exhibited by most magnitude data, including business, manufacturing and construction data, options for post-tabular SDL are limited to cell suppression and controlled tabular adjustment.

As with threshold rules, dominance rules were developed before the advent of computers, and for the case of establishment-based economic data. *Primary disclosure analysis* (identifying the disclosure cells) was performed by hand by individual analysts. The tools were a printed listing of the cells containing the contribution data, and a desk calculator (for addition and percentage calculations). Consequently, dominance rules suffer one weakness: as subjects (companies) can contribute to multiple cells, a proper disclosure analysis would protect company data (e.g., to within p-percent) as cell combinations are formed. This is a daunting task—even today—as in principle one must examine every potentially computable aggregation of cells for disclosure. Similarly, operationalization of an SDL method such as complementary cell suppression to take into account disclosure in cell combinations is challenging. But, at least, the *p*-percent rule and linear sensitivity measures do enable quantification of disclosure risk.

From a quality standpoint, controlled tabular adjustment (*CTA*) is the superior choice for magnitude data. CTA enables release of fully populated tables meeting the protection requirements imposed by the sensitivity measure. Local quality can be assured by imposing constraints on changes to nondisclosure cells. Global quality can be ensured by preserving distributional parameters (Cox et al. 2004) or shape (Cox, Orelien and Shah 2006).

Concluding Comments

The principal challenge facing modern disclosure limitation is to develop SDL methods that (1) sufficiently reduce disclosure risk and (2) preserve key quality characteristics of original data, including usability and reliability of inference. To do so, meaningful measures of disclosure risk and data quality need to be developed and integrated. This promises to be challenging, based upon simple considerations developed here for measuring and reducing disclosure risk in tabular data.

REFERENCES

Cox, LH (1981), "Linear Sensitivity Measures in Statistical Disclosure Control," *Journal of Statistical Planning and Inference* **5**, 153-164.

Cox, LH (1987),"A Constructive Procedure for Unbiased Controlled Rounding," *Journal of the American Statistical Association* **82**, 520-524.

Cox, LH (2007), “Contingency Tables of Network Type: Models, Markov Basis and Applications,” *Statistica Sinica* **17**, 1371-1393.

Cox, LH (2008), “A Data Quality and Data Confidentiality Assessment of Complementary Cell Suppression,” in: **Privacy in Statistical Data Bases 2008, Lecture Notes in Computer Science 5262** (J. Domingo-Ferrer and Y. Saygin, eds.), Heidelberg: Springer-Verlag, 13-23.

Cox, LH (2009a), “Vulnerability of Complementary Cell Suppression to Intruder Attack,” *Journal of Privacy and Confidentiality* **1**(2), 235-251. http://jpc.stat.cmu.edu

Cox, LH (2009b), “An Examination of Two Methods for Controlled Tabular Adjustment of Tabular Data That Preserve Data Quality,” in: **Work Session on Statistical Data Confidentiality, Manchester, 17 to 19 December 2005, Eurostat Methodologies and Working Papers,** Luxembourg: European Communities, 158-167.

Cox, LH and LR Ernst (1982), "Controlled Rounding," *INFOR: Canadian Journal of Operations Research and Information Processing* **20**, 423-432.

Cox, LH et al. (2004), “Balancing Quality and Confidentiality for Multivariate Tabular Data,” in: **Privacy in Statistical Databases, Lecture Notes in Computer Science 3050** (J. Domingo-Ferrer and V. Torra, eds.), Berlin: Springer-Verlag, 87-98.

Cox, LH and JA George (1989), "Controlled Rounding for Tables with Subtotals," *Annals of Operations Research* **20**, 141-157.

Cox, LH**,** AF Karr and S Kinney (2011), “Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think But Not How to Act (with discussion),” *International Statistical Review*, to appear.

Cox, LH and JJ Kim (2006), “Effects of Rounding on the Quality and Confidentiality of Statistical Data,” in: **Privacy and Statistical Data Bases 2006, Lecture Notes in Computer Science 4302** (J. Domingo-Ferrer and L. Franconi, eds.), Heidelberg: Springer-Verlag, 48-56.

Cox, LH**,** JG Orelien and BV Shah (2006), “A Method for Preserving Statistical Distributions Subject to Controlled Tabular Adjustment,” in: **Privacy and Statistical Data Bases 2006, Lecture Notes in Computer Science 4302 (**J. Domingo-Ferrer and L. Franconi, eds.), Heidelberg: Springer-Verlag, 1-11.

deLoera, J and S Onn (2004), “All Rational Polytopes Are Transportation Polytopes and All Polytopal Integer Sets are Contingency Tables,” **Lecture Notes in Computer Science 3064**, Heidelberg: Springer-Verlag, 338-351.