

## DISCUSSION OF PAPERS BY ROBERTSON/SCHIOPU-KRATINA AND BOUDREAU

David A. Binder<sup>1</sup>

### 1. OVERVIEW OF THE ISSUES

Statistics Canada must, by law, ensure the protection of respondents' confidentiality. This protection is also important when soliciting cooperation from the respondents. In the Statistics Act, it is stated:

No person, other than persons employed or deemed to be employed under the Statistics Act, shall disclose or knowingly cause to be disclosed, by any means, any information obtained under the Statistics Act in such a manner that it is possible from the disclosure to relate particulars obtained from any individual return to any identifiable individual person, business or organization.

From the technical perspective, the risk of statistical disclosure is difficult to assess. In theory, at least some disclosure occurs whenever one has more information about an individual than before the release of the data. However, if the data release has any value, then disclosure in this broad sense is bound to occur. It is only when an individual's data can be inferred to within a narrow range that controls need to be put in place. We see that some judgment will be needed here. Some questions that immediately arise are:

1. Is the "Individual" for whom the data need protection well defined? This question is natural when the universe is hierarchical, such as is the case for most business surveys. It is also a problem for household survey data where household members, families, or other groups (e.g., persons illegally living in a particular condemned dwelling) may need protection.
2. How does one determine when a range is sufficiently narrow so as to be sensitive? Any definition for sensitivity of the data will lead to a situation where a small change in the data can change the status of the information from being non-sensitive to being sensitive, or vice versa. This may not be intuitively reasonable, so that exceptions to the rule may be necessary,

especially for repeated surveys.

3. Even in the extreme case where exact disclosure may occur, it is not always possible to determine this. This is because the impact of the release of data will depend on how much prior information the intruder has about the individual. For example, the intruder may have some information about the individual which is not widely known and which is so rare as to effectively identify the individual.
4. Some characteristics are more widely known for some individuals than for others. This would be true particularly for public/sports figures. Therefore, it may not be possible to use the same procedures for all individuals.

Some attempts have been made in the literature to quantify the risk of disclosure by estimating certain probabilities and thereby assessing the risk that may occur. Boudreau has considered the probability of uniqueness, but this is only part of the story. Even when an individual is not unique, his value may be known to within a narrow range if all individuals with the same value for the key variables have a similar value. As well, the structure of the population and the information available may be such that an individual's value can be inferred to within a narrow range using methods of statistical inference. Papers by Duncan and by Lambert, for example, address this issue. Whereas, the structure of this approach using statistical inference may be useful for guiding us on the best methods, such probability methods have not yet been useful as the definitive method for disclosure control.

Because of all the above considerations, the methods for disclosure control have all been heuristic, at least to some extent. For many methods, mathematical formulations have been developed to ensure that certain conditions are satisfied and these have led to much of the technical research. However, it is important to keep in mind that in the absence of universally acceptable measures of disclosure risk, such mathematical

---

1

David A. Binder, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6, BINDDAV@statcan.ca

formulations do not provide the definitive solution answer to the problems being addressed. This is why it is important to understand the properties of such procedures and to allow common sense to override the mathematical results when the results do not coincide with intuition.

Other considerations must also be taken into account. Although a particular data release may be perfectly acceptable on legal grounds, it may be unacceptable because of the negative impact it may have on respondent relations. On the other hand, users would be justifiably annoyed if we do not release data that is can be legally released. This implies that the nontechnical issues may override the technical ones. It is often said that the perception of confidentiality is at least as important as the fact. Therefore, release of data which does not technically violate any rules may still give the appearance of violating the confidentiality requirements. This situation must be recognized and addressed. This problem is becoming more important as we move toward release of richer microdata bases.

## 2. SUMMARY OF DISCLOSURE CONTROL METHODS

Methods to control the disclosure risk of data generally depend on the nature and type of the data. Magnitude data are usually released through tabulations where each cell gives the total for the variable. This is usually cross-classified by various types of classifications, such as geography, industry, size, age groups, etc. For frequency data, the tabulations are counts of the occurrence of a particular characteristic. Finally, release of public use microdata files needs specialized measures to control the disclosure risk.

For magnitude data, some of the measures to limit disclosure are to avoid releasing information with small numbers of respondents, and to check for dominance of the cell by a few large contributors. The case of dominance rules is discussed in detail in the Robertson and Schiopu-Kratina paper. One of the common methods of control is to suppress sensitive cells. However, suppressing only sensitive cells leads to the need to suppress other cells since marginal totals are available. This is known as *complementary suppression* and is discussed at length in the Robertson and Schiopu-Kratina paper. Other methods of control include reducing the size of the table, rounding (conventional, random, or controlled), and random perturbation by perturbing the output directly or the data themselves.

The CONFID system developed at Statistics

Canada computes the range for a cell that is implied by the suppression pattern. This avoids the difficulty of numerically computing the rank of a matrix to determine whether any cell value can be derived through a set of linear equations. The user can also determine whether the implied range may be too narrow. However, these implied ranges are not published, even though they can be mathematically derived directly from the table. It is generally felt that the methods used in CONFID are state-of-the-art.

Some of the current research issues that need to be addressed are how to protect against inferential disclosure resulting from the added knowledge of similar data released over time in a repeating survey or of related data from other sources.

Since issues associated with frequency data are not addressed by either of the papers of this session, I will not comment further on these here.

Disclosure control methods for public use microdata files must address a number of types of disclosure. Boudreau discusses some of the mathematical issues associated with uniqueness of individuals on the file for sample surveys. There are, of course, other forms of disclosure such as by response knowledge (i.e., intruder knows person is in sample), and spontaneous recognition (e.g., famous people). Conceptually, it is required to ensure that the probability that an intruder can identify at least one respondent is small, but this probability is difficult to evaluate in general.

Methods to control the disclosure risk are not addressed in the two papers. Some of them include data reduction, sampling, top coding and bottom coding, suppression, random perturbation, data swapping and microaggregation.

Additional difficulties can arise with longitudinal microdata files, since it is often necessary to specify how variables will be suppressed or masked for survey waves that will take place in the future.

## 3. CONCLUSION

In summary, we see that there are many research issues that arise from the requirement to limit the disclosure risk. Since Statistics Canada would like to ensure that as much of its data are as widely disseminated as possible, a better understanding of these disclosure risks can help fulfill this objective. The two papers presented in this session make some headway in this direction.