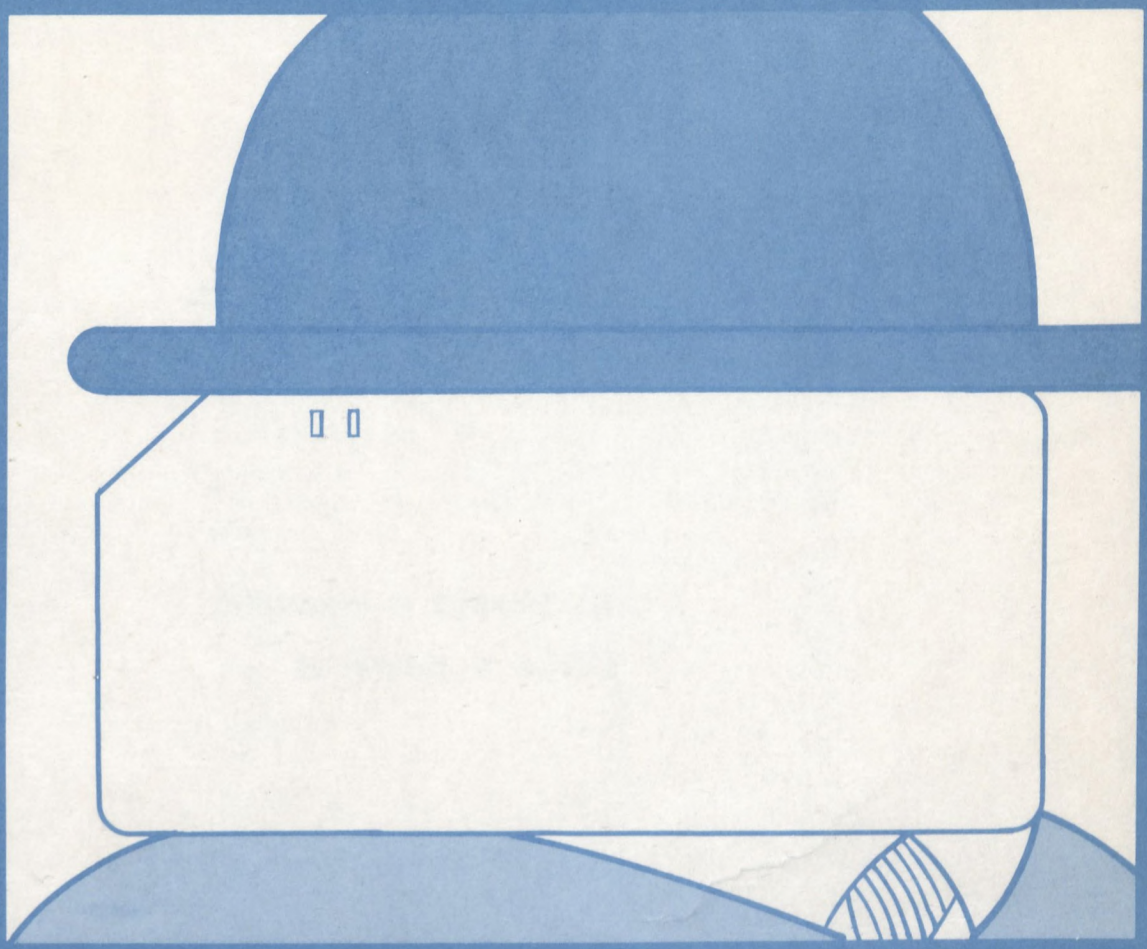
STATISTICAL DATA BANKS AND THEIR EFFECTS ON PRIVACY

H.S. GELLMAN



5

A study by the Privacy and Computer Task Force

STATISTICAL DATA BANKS AND THEIR EFFECTS ON PRIVACY

A STUDY FOR THE
PRIVACY AND COMPUTERS TASK FORCE

DEPARTMENT OF COMMUNICATIONS

DEPARTMENT OF JUSTICE

H.S. Gellman

This report was prepared for the Privacy and Computers Task Force, an inquiry sponsored by the Departments of Communications and Justice, and should not be construed as representing the views of any department or of the Federal Government. The views expressed herein are exclusively those of the authors, and no inference of any commitment for future action by any department or by the Federal Government should be taken from any recommendations contained herein.

This report is to be considered as a background working paper and no effort has been made to edit it for uniformity of terminology with other studies.

TABLE OF CONTENTS

STATISTICAL DATA BANKS AND THEIR EFFECTS ON PRIVACY

INTRODUCTION

As an industrialized country, Canada poses large and complex problems to its managers in industry and government. Solving these problems by trial and error methods is undesirable because it can aggravate the situations and make the problems worse.

Most managers have come to recognize that even though human decisions require intuition, common sense and good judgement, the more facts available on which to base decisions, the better the decisions will be. As a result, most large industrial and governmental organizations are consciously trying to develop systematic ways of planning and managing their programs and activities.

Such systematic approaches, of course, require data, and the planning will not be effective unless the data is appropriate, accurate and timely. Thus, in all industrialized countries there is a continuing need for comprehensive statistical programs.

In recent years, all levels of government in Canada have been trying to develop more detailed plans to deal with particular regional problems. The achievement of these objectives has been aided by the availability of powerful computer systems to process the resultant large sets of complex data.

In the past, governmental statistics bureaus have carefully screened their publications to ensure against disclosure of confidential information about individual respondents. This has always been regarded as an essential practice in order to retain the confidence of the public. Such screening to protect confidentiality has never been easy, but the task is now made much more difficult because computers make it

feasible to handle statistical information in finer subdivided form. This destroys some of the inherent confidentiality provided by the aggregation of data. In other words, inadvertent disclosure of confidential information can now occur through the publication of too much detail, making it possible for the reader to identify information about an individual respondent.

A related problem comes from the increase in social science research that has taken place in recent years. This has created a demand for receiving the statistical data in a sufficiently disaggregated form to enable the research groups to conduct their own analyses. Such disaggregation increases the risk of disclosure of confidential information.

Market research activities conducted by commercial organizations constitute another potential threat to privacy. Computers can help market research groups increase the amount and detail of the data they collect and analyze. This increased capability can lead them to collect more information about more people, thereby increasing the risk that confidential information might be disclosed.

The role of market research and attitudinal surveys (such as the Gallup Poll) is constantly increasing in our society. Market research firms have no vested interest in the privacy of the individuals or groups they study. They are more concerned about protecting the privacy of their clients. Moreover, market research firms have no general codes of ethics or practice.

The advent of computers has helped to publicize the growing public and professional concern that data collected on individuals might be misused. There are fears that statistical and other data banks will, because of the bureaucratic tendencies of modern society, create massive dossiers for each person.

Recent expressions of concern about statistical files have been directed at governmental organizations. For example, in Canada, the United States and Britain, there were public protests about the Census. The public apparently was worried about the invasion of personal privacy because of some questions that were considered to be too personal and because of the possible disclosure of confidential information.

In the United States these public protests led to the establishment of the Decennial Census Review Committee. This Committee submitted a report [1] in July, 1971 to the U.S. Secretary of Commerce, containing a number of recommendations for preserving privacy. Some parts of that report will be referred to in later sections of this paper.

The purpose of this paper is to identify some of the potential invasions of privacy that can occur through the preparation and use of statistical information. Some possible safeguards will also be suggested, together with comments about their probable effectiveness.

PRIVACY, STATISTICAL CONFIDENTIALITY AND DATA SECURITY

It is important to distinguish between the concepts of privacy and confidentiality. In a recent report [2], a U.S. Government Panel on Privacy dealt with this in a useful way. It states that:

> "...there is a good deal of confusion surrounding the concepts of confidentiality and privacy. Not only is there confusion about the meaning of each separate concept, there is also a tendency to refer to one when the other is intended. Faced with this problem of distinguishing and defining, we decided to assign workable, generally acceptable definitions that seemed to meet our needs.
>
> The dictionary defines privacy as 'the state of being private'. The word 'private' has a multitude of definitions...but central to most of them is the concept of the personal as opposed to the public. As applied in the context of the government seeking information, the right to privacy therefore may be defined as the individual's right to decide whether and to what extent he will divulge to the government or its representatives his thoughts, his feelings, and the facts of his personal life. It is a right which is essential to the maintenance of human dignity and freedom of self-determination, and whenever a government authority demands information and imposes a penalty, sanction or forfeiture on those who refuse to comply with the demand, it abrogates the right to privacy with respect to the information demanded, and thereby diminishes the freedom of those upon whom the demand is made.
>
> Confidentiality, on the other hand, is a word that denotes a particular status of information. Information held in a confidential status is subject

> to a restriction or series of restrictions on transmission. Since the nature of the restrictions may vary, especially as now used by the federal government agencies, the word 'confidentiality' is one of imprecise meaning. Often, it denotes not only restrictions on transmission, but also restrictions on the purposes for which particular information may be used."

Except where the context indicates otherwise, the word "confidentiality", when used in this paper, describes a status of information in which the capability of the recipient of information to transmit it and use it is subject to restrictions.

Statistics Canada operates under regulations of the Statistics Act which contains the following two basic provisions.

1. The Bureau is authorized to collect information from respondents (and arrange to penalize those who refuse to give the requested information).

2. The Bureau must not disclose information provided by individual respondents.

Section 29 of the Statistics Act of Canada states:

> "Every person who, without lawful excuse,
>
> (a) refuses or neglects to answer, or wilfully answers falsely, any question requisite for obtaining any information sought in respect to the objects of this Act or pertinent thereto that has been asked of him by any person employed or deemed to be employed under this Act, or
>
> (b) refuses or neglects to furnish any information or to fill in to the best of his knowledge and belief any schedule or form that he has been required to fill in, and to return the same when and as required

of him pursuant to this Act, or knowingly gives false or misleading information or practises any other deception thereunder

is, for every such refusal or neglect, or false answer or deception, guilty of an offence and is liable on summary conviction to a fine not exceeding five hundred dollars or to imprisonment for a term not exceeding three months or to both."

Section 30 of the Act states:

"Every person

(a) who, having the custody or charge of any documents or records that are maintained in any department or in any municipal office, corporation, business or organization, from which information sought in respect of the objects of this Act can be obtained or that would aid in the completion or correction thereof, refuses or neglects to grant access thereto to any person authorized for the purpose by the Chief Statistician, or

(b) who otherwise in any way wilfully obstructs or seeks to obstruct any person employed in the execution of any duty under this Act

is guilty of an offence and is liable on summary conviction to a fine not exceeding one thousand dollars or to imprisonment for a term not exceeding six months or to both."

Section 16 of the Statistics Act states:

"(1) Subject to this section and except for the purposes of a prosecution under this Act,

(a) no person, other than a person employed or deemed to be employed under this Act, and sworn under section 6, shall be permitted to examine any identifiable individual return made for the purposes of

this Act; and

(b) no person who has been sworn under section 6 shall disclose or knowingly cause to be disclosed, by any means, any information obtained under this Act in such a manner that it is possible from any such disclosure to relate the particulars obtained from any individual return to any identifiable individual person, business or organization.

(2) The Minister may, by order, authorize

(a) the particulars of any information obtained in the course of administering this Act to be communicated to a statistical agency of a province pursuant to an agreement under section 10; and

(b) the particulars of any information collected jointly with a department or corporation pursuant to an agreement under section 11 to be communicated to the department or corporation that was party to the collecting of the information.

(3) The Chief Statistician may, by order, authorize the following information to be disclosed:

(a) information collected by persons, organizations or departments for their own purposes and communicated to Statistics Canada before or after this section comes into force, but such information when communicated to Statistics Canada shall be subject to the same secrecy requirements to which it was subjected when collected and may only be disclosed by Statistics Canada in the manner and to the extent agreed upon by the collector thereof and the Chief Statistician;

(b) information relating to a person or organization in respect of which disclosure is consented to in writing by the person or organization concerned;

(c) information relating to a business in respect of which disclosure is consented to in writing by the owner for the time being of the business;

(d) information available to the public under any statutory or other law;

(e) information relating to any hospital, mental institution, library, educational institution, welfare institution or other similar non-commercial institution except particulars arranged in such a manner that it is possible to relate such particulars to any individual patient, inmate or other person in the care of any such institution;

(f) information in the form of an index or list of

(i) the names and locations of individual establishments, firms or businesses,

(ii) the products produced, manufactured, processed, transported, stored, purchased or sold, or the services provided, by individual establishments, firms or businesses in the course of their business, or

(iii) the names and addresses of individual establishments, firms or businesses that are within specific ranges of numbers of employees or persons engaged or constituting the work force; and

(g) information relating to any carrier or public utility."

Section 33 of the Act states:

"Every person who, after taking the oath set out in subsection (1) of section 6,

(a) wilfully discloses or divulges directly or indirectly to any person not entitled under this Act to receive the same any information obtained by him in the course of his employment that might exert an influence upon or affect the market value

of any stocks, bonds or other security or any product or article, or

(b) uses any such information for the purpose of speculating in any stocks, bonds or other security or any product or article

is guilty of an offence and is liable on summary conviction to a fine not exceeding five thousand dollars or to imprisonment for a term not exceeding five years or to both."

These provisions of the Statistics Act of Canada are intended to ensure the following results.

1. The information submitted by persons, businesses, etc. will be used only for statistical purposes.

2. The data that is collected and stored is handled securely. That is, adequate data security procedures are followed.

3. Statistics Canada will maintain a continuous scrutiny of its publications to prevent the deduction of information about particular respondents. That is, the Bureau will adopt procedures to prevent inadvertent disclosure of statistical information.

Apart from the legal need to protect the confidentiality of the information it receives, the Bureau recognizes that such action is vital. The Bureau depends on the trust of the public and cannot afford to lose this trust. It can force the public to submit information, but the accuracy of that information will depend on the public's confidence in the Bureau.

An indication of the degree of trust the public has in Statistics Canada is the fact that some Canadian farmers listed marijuana in the other crops category in the Census questionnaire which asked how they used their land last summer, and listed marijuana sales in the etcetera category

under sources of income for 1970. These farmers know that the censustaker is bound by law not to tell who and where they are — especially not to tell the RCMP [3].

Statistics Canada is interested in information about statistical populations, not individual respondents. The objectives of the Bureau, therefore, do not represent any threat to privacy.

With regard to potential breaches of data security, the Bureau must and does remain vigilant. The Bureau throughly indoctrinates its employees to make sure that they understand that information is confidential, not only when it relates to individuals but also when it is in aggregate form [4]. In addition, the Bureau has implemented strict internal security arrangements and continaully tries to improve them.

Computers have introduced new complexities and difficulties regarding data security. For example, it is more difficult to determine when an unauthorized person makes a copy of a computer magnetic tape that contains statistical information. With older models of computers this was easier to prevent because only one job at a time could be processed. With some of the new computer systems, several jobs can be processed simultaneously. This means that several users can share a single computer system and it is necessary to ensure that one user cannot gain access to another's files through his manipulation of the programs and data in the computer.

Data security is not easy to ensure when several users share a single computer system through telecommunication links. Procedures are continually being developed to improve data security in shared computer systems, but the problem has not yet received a completely adequate solution. In his analysis of this situation, Canning [5] has suggested that in most typical shared computer systems that use remote terminals, "it is unwise to put any highly sensitive data on line to the computer".

The risk of accidents will always remain as long as people are involved in handling data. Mistakes can occur that cause copies of confidential data to be mailed out. A computer operator might mislabel a magnetic tape, leading to its being printed and seen by unauthorized persons. Confidential reports could be misfiled and placed in a library that is open to the public. To protect the confidentiality of information, it is essential to exercise vigilance not only by implementing effective data security systems but also by guarding against mistakes.

Statistics Canada gives information to provincial statistical agencies, on the condition that these agencies maintain the same standards of data security as does Statistics Canada. Although the most thorough care is taken by all concerned, it is apparent that the more hands, and the more institutions, through which a particular item of information passes, the greater the danger of a breach of confidentiality at some point along the lengthening chain.

INADVERTENT DIRECT DISCLOSURE

Even if a statistical office has adequate data security procedures, possible breaches of confidentiality could still occur. Disclosure of information can result from the publication practices of the statistical office. The office would not provide confidential information about a fully identified person, establishment or other respondent, but if the office publishes too much detail, the respondent might be inadvertently identified.

Fellegi [6] has dealt in detail with this problem, which he calls "inadvertent direct statistical disclosure". He points out that in the area of economic statistics the major characteristics of the largest respondents (and often even their identity) is common knowledge. The published statistics typically involve quantities such as production, sales, employment and prices. Since the identity of some of the larger respondents is often common knowledge, care must be taken to avoid identifying what they report. In the words of Fellegi,

> "It is a generally accepted practice to blank out information which is based on fewer than three respondents on the assumption that any two respondents of a particular kind might easily know of each other and hence, if a statistic based on two respondents were published then any one of the two could subtract his own report from the published aggregate and would thus deduce the quantity reported by the other. When there are more than three respondents but one or two respondents account for more than a specified proportion of the aggregate the information is also blanked out. This is obviously necessary in the case of highly skewed distributions where the number of respondents by itself is hardly an appropriate guideline."

In the field of economic statistics, it is also useful for the statistical office to maintain registers of organization units and their relationships. This can help the office to check for possible disclosure, because an aggregate derived from a survey of establishments may satisfy all the requirements described above, but if several of the establishments entering the same tabulation cell belong to the same enterprise, the publication may disclose confidential information about that enterprise. By using a register of business units, the statistical office can check for disclosure at both the establishment and the enterprise level.

Thus, in the field of surveys of business units there are reasonably precise rules to determine whether or not a particular tabulated number represents inadvertent direct disclosure. In the field of socio-demographic statistics, there are less precise rules to work with to prevent direct disclosure. In this area disclosure checking is generally an intuitive process. Most of the published data refers to estimated numbers rather than quantities. The contribution of any one person to a tabulation cell is either zero or one. There is little danger of disclosure as long as the cross-classifications involved do not become so detailed that there are only one or two persons or households of that type. Income statistics receive special treatment. The relatively few people with very high incomes are always "hidden" in a broadly defined income class.

In the case of a tabulation of frequencies from a census, one may argue that there is no violation of confidentiality in a table in which some of the cells contain entries of one, but in which none of the marginal totals are ones. But if another dimension of breakdown is superimposed on that table, then disclosure could occur.

For example, consider a census tabulation that shows a cross-classification of persons by industry and occupation. If one

of the entries in the table is one, but none of the marginal totals are ones, the table may show that there is one person in the textile industry whose occupation is physicist. The reader may recognize the person to whom the entry of one refers. He may say, "Tom Brown is a physicist working in a synthetic textile mill; the table shows that there is one such person, that entry must therefore refer to Tom Brown".

The reader must know in advance something about Tom Brown in order to deduce this information from the table. If the table is now extended to a cross-classification of industry by occupation by salary, at that point the reader may learn Tom Brown's salary. Thus, inadvertent statistical disclosure will occur.

Fellegi [7] has proposed the following definition of inadvertent direct disclosure.

> "Inadvertent direct disclosure in the case of frequency tables based on a census could...be defined as an entry of one in a table, provided that at least one of the corresponding possible marginal totals is also one."

To cope with the problem, Fellegi concludes that:

> "...in the case of tabulated aggregates from census data, one must stop <u>before the level of detail</u> where identification becomes possible, i.e. where one of the entries in the table is based on one, or whatever other specified number of observations. In the case of counts or frequencies one must stop <u>at the level of detail</u> where identification becomes possible. The same criteria might be applied to sample data except that the identity of sample persons should be kept confidential and the level of detail at which disclosure might occur should be considered in relation to the population to which the sample estimates refer."

As a result of the above requirements, Statistics Canada tends to limit the amount of detail in its publications. For example, it might aggregate aluminum production with tin production to protect the Aluminum Company of Canada.

This type of aggregation reduces the utility of the information and there is a constant pressure upon Statistics Canada to be more specific. In contrast, there is little pressure — except at Census time — for the Bureau to be more cautious in order to preserve confidentiality and protect privacy.

RESIDUAL DISCLOSURE

As discussed in the previous section, direct statistical disclosure refers to the case where information about an identifiable individual respondent can be deduced from a tabulation. This is not a trivial problem and its solution depends on the careful application of the quantitative methods described earlier.

An even more complicated problem is that of residual or complementary disclosure. Residual disclosure occurs when a set of tabulations can be manipulated arithmetically to yield, through deduction, information about an identifiable respondent, even though no single tabulation discloses information about an identifiable respondent [8].

For example, residual disclosure could occur if an entry in a table is blanked out but that entry could be deduced from the marginal totals and the other entries in the table. Each time a new tabulation is produced from the same survey data, a new disclosure could occur through the arithmetic manipulation of the set of tabulations. It is necessary, therefore, to devise methods that will prevent such disclosures.

Fellegi has dealt with this problem [9] by developing a mathematical test that involves the calculation and comparison of very large matrices. A detailed description of Fellegi's method is contained in Appendix 1 (page 35).

Unfortunately, the above method is not a practicable solution, even if the largest available computers are used.

A more promising solution is to introduce a minor level of random disturbance into every table. This approach has been tested by Statistics Canada and will probably be adopted by them [10]. Their method is called random rounding.

To apply this method, one would generate a set of random numbers in a computer. This is equivalent to tossing a die or pair of dice. Efficient methods are available for generating random numbers in computers. Most of the methods in current use involve the multiplication of two numbers (one of which is a specially selected constant).

Statistics Canada plans to publish cell entries which are multiples of 3. If the "true" number is divided by three and leaves a remainder of one, they will multiply a random number by a number from a probability distribution to cause two-thirds of the numbers to be rounded down and one-third to be rounded up. If the remainder is 2, two-thirds of the numbers will be rounded up and one-third will be rounded down.

Thus, if a published tabulation cell contains the number 3, the "true" number could be 2, 3 or 4. This "disturbance" prevents residual disclosure because the additions and subtractions will not "balance". This procedure will add minor errors to the tabulation, but these errors will be small compared to other errors that already exist in the tabulation.

If ordinary rounding procedures were used (instead of random rounding), the probability that a number is rounded up is equal to the probability that it is rounded down (instead of the three to two ratios involved in random rounding). Ordinary rounding provides some protection against residual disclosure but random rounding provides more.

In Holland the census bureau plans to use ordinary rounding procedures on its current census and will publish its numbers as multiples of 5. Statistics Canada plans to use random rounding only for special tabulations requested by users (called ad hoc tabulations).

A related "randomized" approach has been advocated by Boruch [11]. He states that:

> "Typically, the researcher attempts to maintain an isomorphic relation between a person's responses on a questionnaire and records of these responses transformed to magnetic tape form. Now, the possibility of data use or misuse is, of course, weakened when data are not reliable for any specific individual record. Frequently, the researcher can afford to undermine deliberately the integrity of a single record but preserve the integrity of the whole, at least with respect to statistical parameters. He can do so by inoculating statistical data files with randomized errors whose properties are known. A large body of literature deals with the problem of adjusting statistical estimates of population parameters, when the observations are subject to known measurement error. The inoculation accomplishes a number of important objectives. First, in the context of public interest in survey research, confusion between administrative records, eavesdropping devices, intelligence systems etc., may be minimized. The controlled unreliability of any individual record is a notion that can be communicated to the public. Second, the likelihood that records will be used in formation of judgements about specific individuals is reduced substantially. One cannot obtain unambiguous information about a specific person, even if identification is, in fact, accomplished."

Using Boruch's approach, to inoculate a statistical table that consists of only zeros or ones (i.e., yes or no answers to questions) one could proceed as follows. With the toss of

one die, if the die shows 1, 2, 3, 4 or 5 we would leave the yes or no answer unchanged. If the die shows 6, we would change yes to no and no to yes for the particular tabulation cell.

This procedure involves a known probability distribution. That is, we know that we will introduce errors in one-sixth of the cases and this knowledge enables us to deduce the "correct" statistical averages and other parameters. Yet, the person who reads the inoculated table cannot know which cells are "true" and which are "false".

Both of the "random" methods described above appear to offer practicable and effective solutions to the problem of residual disclosure.

RECORD LINKAGE

It is desirable and valuable to bring together existing data about the same person that have been collected by different agencies. Such linkages of records are useful for statistical research studies and program evaluations. In addition, the linkage of reports of changes of addresses, marriages, births, etc., among different administrative data files is desired by most citizens in order to reduce the need to report such changes separately to each agency.

There is, however, a danger that protection of confidentiality will be weakened through linkages of data. In many research situations, if name and address files were matched with statistical information files, the total file would comprise an intelligence system. However, if additional coding is used at some separate centre to provide the basis for a double linkage system, such matching can be prevented. For example, Boruch [12] describes a method in which the name and address file is kept separate from the statistical record file. Each individual record in a given statistical data file is assigned a unique (arbitrary) accounting number. Each record in the corresponding name and address file is assigned a different accounting number. A code array of numbers, which match numbers in the first set to the corresponding numbers in the second set, is created. The code linkage can be maintained by a separate organization or various security procedures can be applied.

Despite the availabity of methods to protect confidentiality in record linkages, we do not have adequate standards of practice to ensure that the available precautions will be taken by the owners of data files. As Bachi and Baron have stated:

> "On the one hand inefficient collection of similar data by different government agencies should be

avoided where possible as modern methods of record-linkage enable the wide use of specific items of data, opening new vistas for statistical research to serve the growing needs. On the other hand, statistical agencies must take measures to ensure that the public can rely on the preservation of the confidentiality of the data entrusted to them by respondents over the years" [13].

PERSONAL IDENTIFICATION NUMBERS

Computer filing systems work much better if every record carries a code number; names and addresses are unreliable factors where the matching of different records must take place entirely logically. The computer is too logical for this problem. Human intelligence will make the reasonable assumption that a Mr. A. Gibbings of 12 Thorncliffe Park Drive in one file, is almost certainly the same person who appears elsewhere as Mr. A.G. Gibbins of 12 Thornclyffe Avenue. The computer much prefers to know him consistently as 418-851-218.

Code numbers also help to avoid duplication in the cases where different people have identical names. In addition, the exchange of data among computer systems is less expensive with code numbers because the number of digits in the identification number is considerably less than the number of digits and letters in the name and address. Thus, less space is needed inside the computer to store the identification number and less time is needed to sort the file.

Identification numbers for all citizens have already been introduced in Sweden (1947), Israel (1948), Norway (1964), Finland (1965) and Denmark (1968). Preparations are underway in Argentina, the Benelux countries, the Federal Republic of Germany, Japan, Switzerland, Spain, South Korea, the USA and East Germany [14].

In Denmark the "person number" consists of 10 digits of which the first 6 include information of the person's birthday (in the order: day-month-year), while the last 4 digits are a serial number. The 10th and last digit is a check digit [15].

The Federal Republic of Germany is currently planning to introduce a "personal identification number" comprising 12 digits. The first 6 digits will be for date of birth, the 7th will identify sex, the next 4 will be a serial number and the last will be a check digit [16].

The adoption of personal identification numbers leads to the possibility that we will all be reduced to "mere numbers" and this disturbs some people. It reminds them of the conditions of war, when the necessity of greater control of individual movement and action is accepted as an unfortunate necessity. We in Canada may use credit cards, but identity papers are still taboo.

In Denmark there was not much opposition before the person number system was established. However, after it was implemented there was considerable criticism — especially by the press. The criticism centred on the possibility that person numbers would make it easier to collect data on citizens and that this information might be misused [17].

In the United States, the proposal to include the social security number in the 1970 census was dropped after considerable opposition in Congress and elsewhere [18].

A recent survey conducted by the American Federation of Information Processing Societies and Time Magazine showed that 54% of the respondents believe computers are dehumanizing people and turning them into numbers. 62% are concerned that some large organizations keep information about millions of people. In addition, 53% believe computerized information files might be used to destroy individual freedoms; 58% feel computers will be used in the future to keep people under surveillance [19].

The fear that data banks might create personal dossiers has raised objections to the use of the Social Insurance Number in Canada. In an editorial in July, 1970 [20], the editor of the Canadian Chartered Accountant magazine wrote:

> "When the Canada Pension Plan resulted in the assignment of a Social Insurance Number to virtually every Canadian, those who pointed out that the number could be readily adopted as a permanent

> code in a much broader application were dismissed as obstructionists or alarmists. Yet what would the reaction be if all retail credit grantors decided to use this number (or a common account number) for each customer? In short, some of the population seems mesmerized by the notion that government will not misuse information or the machinery for gathering it. When historical examples are cited to point out the fallacy of such thinking, the responses ranges from 'it can't happen here' to 'it's necessary for progress'! Doubtless, a government bent on complete subjugation of the people could achieve those ends regardless of the country's information system if it exercised sufficient cunning in the early stages. But why, for the short-term 'pay-off' we might gain, should we build a system that would be pure gold in the hands of the ruthless?"

In October, 1970, a data privacy act was proposed by British Columbia M.P. Tom Goode. Mr. Goode is quoted [21] as saying that:

> "Social Insurance Numbers could easily be used to infringe on our privacy. The number can allow dozens of computers to trade information about us in such a way that a complete record of our dealings, activities and associations can be built up."

Mr. Goode was answered in a letter to the editor of Canadian Datasystems by Mr. Balmer [22]. Mr. Balmer suggests that:

> "To deprive [our commercial enterprises] of the use of a unique personal identification code because its use would be abused is tantamount to saying all good citizens should be subject to a dusk-to-dawn curfew because some of them may conduct nefarious activities under cover of

> darkness...Please remember that unless you permit accurate person identification to be made you take the risk of being mistaken for an (unworthy) person."

The use of personal identification numbers could lead to increased data-gathering or to data-concentration. In the words of Professor Miller, (23)

> "The new information technologies seem to have given birth to a new social virus — data mania. Its symptoms are shortness of breath and heart palpitations when contemplating a new computer application. A feeling of possessiveness about information, and a deep resentment toward those who will not yield it."

This condition is particularly true with regard to statistical data which can be handled so efficiently by computers. The danger is that as governments acquire greater ability to handle information, they will demand more information. And as they collect more information, they may also distribute it widely. As the Ontario Law Reform Commission has stated,

> "...It is widely accepted that wherever the government licenses, controls or otherwise regulates economic and social activities for the common good in pursuit of deliberate public policies, then it has the right and need to gather enough relevant data to do this efficiently. Yet, the fact that a large mass of personal data about the people in businesses in the province exists in governmental files does not justify either the collection of more than is necessary to implement these policies, or any disclosure outside of either the government department or ministry or the government as a whole to persons who have some interest in the same data for different reasons. The

government should not become the vehicle for distribution of personal information that it happens to possess simply because it has the right and the need to collect it in the first place" [24].

On the other hand, as Warner has written, [25]

"It can always be argued that to allot numbers to people is in fact not destructive of individuality in any way, and the apologists for such a position point to the proliferation of numbers we already carry — Social Security, National Health, Tax Reference, payroll, Armed Services and so on. To agree on one single number for all systems is no more than rationalization. They can also instance the European nations, where a common identity numbering method was introduced wherever Napoleon went, and has remained an operative system ever since. They can tell us that the recent introduction of person-numbering in Scandinavian countries was accomplished with very little objection from moralists, humanists, or psychologists — even though in all three countries (Norway, Denmark, Sweden) the method of allocating numbers makes it possible to derive the person's age directly from the number.

...in Sweden every individual has a number based on his birth date and area of birth. Numbers are better than names, because the latter are often shared, or even changed over the course of a lifetime. Every Swede or settler in Sweden knows this number as well as his name because he quotes it constantly in every single transaction with the state or increasingly in most transactions with private organizations too. One advantage accrues to the individual: he is not cluttered up with a whole string of code numbers issued to him by

different organizations. His birth number is his, and his alone, and it serves him wherever he goes."

The use of personal identification numbers can make record linkage easier but it does not solve all the problems of file integration. For example, even if two sets of computer files were linked through the use of a personal identification number, in many cases the file formats are not compatible so that the information could not be retrieved automatically without human intervention. To make two separate computer files compatible usually involves a considerable amount of time and money for translation and conversion. Therefore, the availability of a personal identification number produces only marginal economic benefits during record linkage with today's computer systems. On the other hand, if personal identification numbers were adopted in Canada, it is likely that system designers would tend to place more emphasis on file compatibility so that in the future, computer files could be linked more easily.

In Canada, at present, there is no personal identification number that is used by every person. Use of the Social Insurance Number is not mandatory. Its use is mandatory only in the Canada Pension Plan, for unemployment insurance and for income tax purposes. Thus, most employed people are included, but many people in Canada do not have a Social Insurance Number.

The Social Insurance Number index is maintained by the Unemployment Insurance Commission and has 13.5 million numbers on file covering almost all the labour force and a number of special groups such as school children. Plans are underway to automatically update the file as a result of births, marriages, deaths and other changes in the personal status of persons covered in the files.

Some provincial government agencies have been reluctant to adopt the Social Insurance Number for their files. For

example, the Ontario Department of Transport (now part of the Ontario Department of Transportation and Communications) did not adopt the Social Insurance Number for driver licences because the Social Insurance Number does not contain enough intrinsic information. Instead, Ontario adopted its own driver licence number system which contains some information about the driver, such as the date of birth.

Ontario does use the Social Insurance Number as the account number for its medical insurance plan. In this case, however, it was necessary to use a "dummy" first digit to accommodate persons who do not have a Social Insurance Number.

The Social Insurance Number comprises nine digits. The first digit specifies a geographic region and the last is a check digit; the other digits have no significance. In contrast, Sweden uses a ten digit number in which the first six digits refer to date of birth, the next three are for geographic allocation and the last is a check digit. In Sweden the personal identification number is issued at birth so that all people have a number.

If a unique personal identification number were available and used widely in Canada, there could be economic benefits from using these numbers to exchange data among various governmental and commercial organizations, such as credit bureaus, chartered banks, etc. It is reasonable to expect increasing pressure to come from commercial organizations for the adoption of such a number.

The American Bankers Association has been studying the idea of a single identifying number for every individual, and leans toward the use of the Social Security Number, since it is already well established as an identifying number. Most of the 250 organizations surveyed by the ABA, ranging from credit-card companies to hospitals, favoured the single

number rather than a different one for each card [26].

As time goes on and people become more accustomed to having numbers assigned to them, a more favourable climate for the adoption of a personal identification number may develop. In recent years, many Canadians have become accustomed to using all number dialing on their telephones, and many are currently being exposed to the introduction of postal code numbers.

It is possible that a "de facto" personal identification number will develop in Canada, either through an ever-widening use of the Social Insurance Number (despite its limitations) or by indirection, through credit card and bank account numbers. However, it is important to ensure that a personal identification number is not adopted in Canada, directly or indirectly, without a full examination and public debate of its merits and consequences.

LEVELS OF PROTECTION

Senior officials of Statistics Canada should be commended for their constant devotion to the protection of confidentiality. Nevertheless, they recognize that their procedures are not perfect. The Bureau publishes more than 140,000 series, so it is difficult to guarantee that no element of disclosure will occur [27]. To this point in time, no known examples of disclosure exist.

The Bureau also recognizes that some disclosures can occur during the enumeration process. It concedes the difficulty of obtaining enumerators of sufficient skill and quality because of the large number of personnel required, the relatively low pay, and the intermittent nature of the work.

It is probably safe to say that if all groups who use statistics about people were to adopt the standards used by Statistics Canada, the Canadian public would have little need to worry about the invasion of its privacy from these sources. At least one research group shares this view. The Institute for Behavioural Research of York University has adopted a code of ethics for its researchers and is using safeguards of confidentiality patterned after those developed by Statistics Canada [28]. The Institute does not have the resources to develop some of the new techniques such as those developed by Statistics Canada to prevent residual disclosure. However, the Institute does not find the level of standards of Statistics Canada to be too high, or difficult to apply.

Unfortunately, this example is not typical. As reported by McPhail [29],

> "...less than fifty percent of the departments and universities have or plan to have ethics committees and...there is such a diversity of

opinions concerning the entire topic area that imminent solutions are unlikely."

It is probably reasonable to assume that the standards of protection of confidentiality are even lower in the area of commercial market research than they are in social research conducted by universities.

While it is true that most market research information is collected from people on a voluntary basis, it is not unusual for the information to be used subsequently by other groups without the respondent's knowledge. For example, a business organization might receive a questionnaire from a market research organization or a firm of auditors or lawyers who say they are conducting a survey on behalf of one of their clients. Many people tend to answer such questions. Yet, if they knew who the client was, they might be reluctant to respond.

In addition to the hazard of this information falling into the hands of unintended recipients, there is also the problem that no code of ethics exists in the area of market research. If it will be difficult to obtain codes of ethics in universities, it will probably be even more difficult to obtain them in commercial market research organizations.

Thus, there is a general tendency for a lowering of standards for preserving confidentiality as we move from governmental organizations to commercial organizations. In fact, in the case of governmental organizations there is probably a rising standard as the provinces adopt legislation to permit them to receive data from Statistics Canada. Statistics Canada cannot provide this data to a province unless the provincial statutory safeguards are at least as stringent as those of the Statistics Act of Canada.

CONCLUSIONS

1. To prevent disclosure of confidential information, agencies that handle statistical files will need to exercise care and vigilance in the areas of data security (to prevent actual theft of information), direct statistical disclosure and residual disclosure.

2. The methods described in the body of this paper can be effective in preventing direct and residual disclosure. The provision of adequate data security is more difficult to ensure because completely adequate methods have not yet been developed. This is particularly true in the case of time-shared systems.

3. It is important to recognize that even if technical safeguards are developed, they will not be effective unless the people involved make good use of them. We need to supervise the people who handle the confidential information — not the tools.

4. In the case of Statistics Canada, we have a reasonably good model for other agencies to emulate. It may, however, be necessary to enforce similar behaviour in other governmental and commercial agencies through some forms of regulation. For example, it may be desirable and feasible to establish licensing systems for data processing personnel, credit bureaus, social science research workers, market research workers, etc. It would, of course, be preferable if these groups adopted and used strong codes of ethics, but this is not likely to occur voluntarily.

   Some professional social science researchers might take offence at being subjected to a licensing procedure. However, medical practitioners can lose their licences

if they behave unethically, so it is reasonable to revoke the licence of a social scientist who divulges confidential information. A similar argument applies to data processing personnel and commercial organizations that handle confidential information. Government employees should not be exempt, just as they are not exempt from requiring drivers' licences.

5. The protection of privacy will usually involve increased costs or reduced efficiencies. Design engineers know that safety mechanisms cost money [30]. But most citizens realize that the cost of protection is usually small compared to the cost of the bad outcome.

6. It is important to balance privacy and efficiency. We should not let ourselves get carried to either extreme. For example, Statistics Canada retains microfilm copies of the original census data, primarily to assist in the certification of a person's age, citizenship and relationship to the household. In a recent U.S. report of the Decennial Census Review Committee [31], the Committee recommends that on the microfilm records, the name should be separated from the main body of information and should be attached only to a limited number of items needed for certification of a person's age, etc.

   In Statistics Canada, the senior officials recognize this problem but a reasonable question is: should a similar separation of these records be made in the Bureau, or is the issue of minor importance?

7. We should not make data linkages unduly difficult to achieve. But we should develop standards to ensure that personal identifiers are controlled by proper agents and are removed from the statistical files after the linkages have been completed.

8. It is difficult to estimate the probability that in Canada the Social Insurance Number will become the de facto personal identification number. However, no personal identification number should be adopted in Canada without a full examination and public debate of its potential benefits and adverse consequences.

9. In view of the high standards developed by Statistics Canada in protecting confidentiality, perhaps the Bureau should be given formal recognition of its accomplishments by establishing it as a leader to develop improved protective procedures in the future. It is almost certain that today's protective methods will not be adequate for long. No sooner will we develop safeguards against one form of disclosure than other, more complex forms will appear. The Bureau need not do this work alone, but it should play a leading role.

10. Every possible method should be used to encourage other government (and private) agencies to adopt the methods used by Statistics Canada in protecting confidential information. However, care should be taken to maintain a reasonable balance. For example, when the new Statistics Act was passed in February, 1971, the Bureau swore in all its employees again and made sure that each employee understood the obligations involved in the job. But the Bureau then went on to consider the fingerprinting and photographing of its 4,000 regular employees for "security" purposes. This brought a protest from a group of employees [32].

Perhaps this is an example of a case where people are trying to achieve perfection. No human can be perfect, nor can institutions or systems designed by humans. The pursuit of excellence should not be disparaged, but if we are not moderate, we run the risk of failing to achieve a good solution because we are not willing to settle for less than a perfect solution. Here, "perfect" is the enemy of "good".

APPENDIX 1

Checking for residual disclosure (counts or aggregates)

A theorem

Each tabulation cell involving counts or aggregates can be conceived of as corresponding to a set of respondents. Clearly, when another estimate of an aggregate, not previously published, is deduced through arithmetic manipulations, it also corresponds to a set of respondents. It only takes a moment of thinking to realize that this set must be an intersection (or union of intersections) of some of the sets corresponding to the previously published data.

In order to check for residual disclosure, it is therefore necessary to consider all the sets of respondents corresponding to published aggregates (publication sets), take all the possible set intersections and unions and answer two questions for each of the resulting sets: would it be i.d.d. if the corresponding count or aggregate was published?; can it be isolated through an arithmetic manipulation of the published aggregates? If both questions are answered in the affirmative for any of these new sets then residual disclosure occurs. Moreover, it is obvious that from the published counts (or aggregates) another count (aggregate) can only be deduced through linear combinations.

All the new sets are unions of elementary intersection sets, where the elementary intersection sets can be defined as the smallest mutually exclusive, non-overlapping sets which can be created through the operation of intersections and complementing from the publication sets.

If a careful account is kept of all the respondents and the published tabulation cells they enter, then the elementary intersection sets referred to above are (conceptually) easily identified: it involves the identification of those respondents who enter a particular two, three, four, etc. publication sets. It is also (conceptually) easy, given a precise definition of i.d.d., to answer for any of the elementary intersection sets (or any union of elementary intersection sets) whether the corresponding aggregate would be an inadvertent direct disclosure. It remains to answer the second question, however: can the corresponding (unpublished) aggregate be deduced from the published aggregates?

Let us consider an example. In Figure 1 below three publication sets are shown.

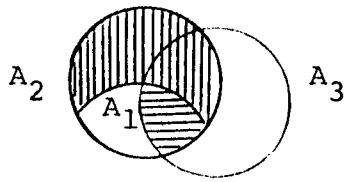


Figure 1

The union of the publication sets A_1, A_2 , A_3 breaks into five elementary intersection sets. In order to determine whether a particular elementary intersection set has a corresponding aggregate which would be i.d.d., we can argue along the following lines, illustrating the argument in terms of the example above.

Suppose that we want to check whether the set which is the intersection of A_1 , A_2 and A_3 would result in residual disclosure. Then we are looking for a linear combination of the tabulation values corresponding to A_1, A_2 and A_3 (say t_1, t_2 and t_3) such that there is a linear combination

$$a_1t_1+a_2t_2+a_3t_3=b_o$$

which is precisely equal to the tabulation value corresponding to the intersection of A_1, A_2 and A_3 , i.e. which is equal to the sum of the values associated with the respondents in that set (the horizontally shaded set). Looking at it differently, consider a respondent in the intersection in question and let the value associated with that respondent be denoted by x. Then the value x is included in all three of the totals t_1, t_2 and t_3. For x to be counted in the linear combination precisely once, it is necessary and sufficient for

$$a_1+a_2+a_3 = 1$$

to hold. Consider now a respondent in, say, the elementary intersection set which is included in A_2 and A_3 but not in A_1. Let the value associated with this respondent be denoted by y. Since this respondent is not in the intersection of all three of the sets A_1, A_2 and A_3, the corresponding value, y, should not be counted in the total b_o. But y is included in t_2 and t_3 (and not in t_1) so for y to cancel out of the total b_o it is necessary that

$$a_2+a_3 = 0$$

should hold.

The pattern should now be clear. There is one equation corresponding to each of the elementary intersection sets. All the equations are in terms of the unknown a_1, a_2 and a_3, each corresponding to a tabulation set. The coefficients in an equation corresponding to an elementary intersection set are determined as follows: those of the unknowns whose corresponding tabulation sets contain the intersection set in question have coefficients equal to one, the other coefficients being zero. The right hand sides of all the equations are zero, except for the one equation which corresponds to the intersection set which we want to test for residual disclosure: the right-hand side of that question being one. The question of whether or not the total b_o is residual disclosure is determined by (a) whether the system of linear equations just described has a solution and (b) whether or not the publication of b would be i.d.d.

In the particular example, we would get the following set of equations:

$$\begin{aligned} a_2 &= 0 \\ a_2 + a_3 &= 0 \\ a_3 &= 0 \\ a_1 + a_2 &= 0 \\ a_1 + a_2 + a_3 &= 1 \end{aligned} \qquad (1)$$

Clearly, these equations do not have a solution, hence no residual disclosure occurs corresponding to the horizontally shaded set (in fact, by permiting the location of zeros and one on the right hand side, it is easy to see that none of the resulting systems of equations have a solution, so there is no residual disclosure corresponding to any of the elementary intersection sets).

Suppose we want to check for residual disclosure a union of elementary intersection sets, say those in the vertically shaded area. It is easy to see, arguing along the same lines as above, that in order to be able to deduce the value of the statistic corresponding to this set it is necessary and sufficient that the set of equations, which is obtained from (1) above by setting the right-hand sides of the first and second equations equal to one and all the other right-hand sides equal to zero, have a solution. The first and second equations are, of course, those corresponding to the elementary intersection sets contained in the vertically shaded area of Figure 1. We get

STATISTICAL DATA BANKS AND THEIR EFFECTS ON PRIVACY

A STUDY FOR THE

PRIVACY AND COMPUTERS TASK FORCE

DEPARTMENT OF COMMUNICATIONS

DEPARTMENT OF JUSTICE

H.S. Gellman

$$
\begin{aligned}
a_2 &= 1 \\
a_2 + a_3 &= 1 \\
a_3 &= 0 \\
a_1 + a_2 &= 0 \\
a_1 + a_2 + a_3 &= 0
\end{aligned}
$$

This set of equations has a solution, namely $a_1=-1$, $a_2=1$, $a_3=0$. Clearly the value of the aggregate corresponding to the vertically shaded area can be deduced from A_1, A_2 and A_3 (in fact, A_3 is not even needed). Hence the publication of the statistics corresponding to A_1, A_2 and A_3 is residual disclosure provided that the statistic corresponding to the vertically shaded area would, if published, be i.d.d.

More generally, suppose that the publication sets (the sets corresponding to all the previous publication cells) are $A_1, A_2, \ldots, A_k$. Let the corresponding elementary intersection sets be $B_1, B_2, \ldots, B_m$. We assume a sequence of testing for residual disclosure which involves first testing for residual disclosure corresponding to individual elementary intersection sets, then corresponding to unions of two elementary intersection sets, then three, four, etc. In order to test for residual disclosure corresponding to the union of the sets $B_{n+1}, B_{n+2}, \ldots, R_m$ we consider the set of linear equations (with coefficients all equal to zero or one) obtained as indicated below. The sequencing of the sets B_i is, of course, purely for convenience.

Consider the matrix M with elements u_{ij} where

$$
\begin{aligned}
u_{ij} &= 1 \quad \text{if } B_i \subset A_j \\
&= 0 \quad \text{otherwise.}
\end{aligned}
$$

Write down the system of linear equations

$$\underline{Ma} = \underline{c} \qquad (2)$$

where $\underline{a}$ is the column vector of unknowns $(a_1, a_2, \ldots, a_k)$ and $\underline{e}$ is the column vector all of whose elements are zero except the (m-n) (last) elements which are equal to 1.

We can now state our general theorem:

Theorem 1: Residual disclosure occurs corresponding to the union of the sets $B_{n+1}, B_{n+2}, \ldots, B_m$ if and only if the system of linear equations (2) has a solution and if the aggregate corresponding to $B_{n+1}, B_{n+2}, \ldots, B_m$ is a direct disclosure.

The proof of the theorem follows the argument outlined in connection with the example in Figure 1.

In order to determine whether or not the system of equations given in (2) has a solution, we need another theorem (proved in the appendix).

Theorem 2: Consider the matrix N obtained from M by omitting the last (m-n) rows of the latter. Then the system of linear equations given by (2) has a solution if and only if the following three conditions hold:

1. the rank of M equals the rank of N plus one;

2. the addition to N of any of the last (m-n) rows would increase its ranks;

3. there is among the last (m-n) rows one row such that all the other of the last (m-n) rows are equal to it plus a linear combination of rows of N.

Theorem 2 provides a necessary and sufficient condition for the solution of equation (2), i.e. for the deduceability of the statistic corresponding to the union of B_{n+1}, B_{n+2}, ..., B_m from A_1, A_2, ..., A_k. Even if it can be deduced, however, the statistic may not be a disclosure.

The following theorem now immediately follows from theorem 2:

Theorem 3: The publication of counts or aggregates which define the columns of M is residual disclosure if and only if the following three conditions hold:

1. there is a set of rows of M say the last (m-n), whose omission reduces the rank of M by one but the omission of all of these rows is required to reduce the rank of M by one;

2. the count or aggregate corresponding to the union of the elementary intersection sets defined by these rows would, if published, be i.d.d.;

3. one of the last (m-n)rows of M is such that all the others are equal to it plus a linear combination of the first n rows M.

The proof of this immediately follows from that of theorem 2 since the first and last conditions of theorem 3 are equivalent to theorem 2 and provide conditions for the deduceability of the statistic corresponding to the union of $B_{n+1}, \ldots, B_m$. Condition 2 affirms that this statistic is a disclosure.

A few notes are in order in connection with these theorems.

1. The existence of residual disclosure can, through the use of these theorems, be tested in a precise and unambiguous way, provided that there is a precise and unambiguous definition of direct disclosure.

2. Whether or not all the publication sets or only some of them are considered, the corresponding elementary intersection sets, by definition, are not empty. The horizontally shaded set, for example, contains all respondents who enter all three of the publication sets A_1, A_2 and A_3. Empty intersection sets are not considered, i.e. the equations corresponding to them are irrelevant, hence the rows corresponding to them in the matrix M should be omitted. This is an important consideration since it materially alters the considerations with respect to residual disclosure.

For example, if all the respondents in A_3 were in the horizontally shaded area of Figure 1, then the other two elementary intersection sets would be empty. The equations (1) now would appear as follows:

$$\begin{aligned} a_2 &= 0 \\ a_1 + a_2 &= 0 \\ a_1 + a_2 + a_3 &= 1 \end{aligned}$$

Now these equations can be solved; actually in whichever row the one on the right-hand side is placed, the resulting equations can be solved. So if the aggregate corresponding to any of the elementary intersection sets would be a disclosure if published, then the publication of A_1, A_2, and A_3 would be a residual disclosure. The difference between the conclusion of this paragraph and the earlier discussion of equations (1) is that there we implicitly assumed all the elementary intersection sets to be non-empty.

3. The procedure of testing for residual disclosure would seem to involve applying theorem 3 to all elementary intersection sets, then to all possible unions of two of them, etc., stopping when the first residual disclosure is encountered (and deducing, of course, that the publication of all of the publication sets A_1, A_2, ...,A_k is illegal). In fact there are some short-cuts. Clearly, if some of the elementary intersection sets would not be i.d.d. even if published, then there is no point in testing for residual disclosure with respect to them. In fact, if i.d.d. is tied to the publication of data relating to fewer than a specified number of respondents, then only those elementary intersection sets and unions of such sets are worth testing which contain fewer than the specified number of respondents.

If the specified number referred to above is equal to two, then the testing need to be carried out only with respect to those elementary intersection sets which contain one respondent (no unions of elementary intersection sets need be considered since the union of two sets, none of which is empty, would contain two or more respondents). In this case theorem 3 assumes a simpler form:

Theorem 3a: The publication of counts or aggregates which define the columns of M is residual disclosure if and only if the following two conditions hold:

1. there is a row of M whose omission reduces the rank of M by one;

2. the count or aggregate corresponding to the elementary intersection set defined by this row would, if published, be i.d.d.

In this case theorem 3b is relevant (its proof follows immediately from that of 3a and is given in the appendix).

Theorem 3b: There is no residual disclosure corresponding to any of the elementary intersection sets if every row of M is linearly dependent on the other rows of M. If a particular row of M is linearly independent of the other rows then there is residual disclosure provided that the aggregate corresponding to this row is a direct disclosure.

Proof of Theorem 2

Suppose that the matrix N obtained by omitting the last (m-n) rows of M, has a rank which is equal to the rank of M minus one; i.e. if r is the rank of N then r+1 is the rank of M. Suppose also that the row specified in condition 3 of theorem 2 is the last row of M.

According to the assumption above and condition 2 of theorem 2, N taken together with the last row of M is of rank r+1 hence it has a square submatrix of order r+1. One of the rows of this submatrix is the last row of M since otherwise N would have rank r+1, contrary to our assumption. Without loss of generality, it may be assumed that this submatrix of order r+1 has as its columns the first r+1 columns of M and as its rows the first r rows of M together with the last. The first r rows of M are therefore linearly independent; since they are also rows of N, it follows that all other rows of N can be expressed as their linear combinations (otherwise the rank of N would be greater than r). Denoting the row vectors of M (and N) by $\underline{u}_i$, there exists therefore scalars λ_{ij} such that

$$\underline{u}_i = \lambda_{i1}\,\underline{u}_1 + \lambda_{i2}\,\underline{u}_2 + \ldots + \lambda_{ir}\,\underline{u}_r \qquad i=r+1,r+2,\ldots,n \qquad (3)$$

Consider now the following system of equations:

$$\begin{aligned}
&u_{11}a_1+u_{12}a_2+ \ldots + u_{1r}a_r + u_{1r+1}\,a_{r+1} = 0\\
&u_{21}a_1+u_{22}a_2+ \ldots + u_{2r}a_r + u_{2r+1}\,a_{r+1} = 0\\
&\qquad\qquad \ldots\\
&u_{r1}a_1+u_{r2}a_2+ \ldots + u_{rr}a_r + u_{rr+1}\,a_{r+1} = 0\\
&u_{m1}a_1+u_{m2}a_2+ \ldots + u_{mr}a_r + u_{mr+1}\,a_{r+1} = 1
\end{aligned} \qquad (4)$$

Since the determinant of this system is not zero, it has a unique solution:

$$a_1=a_1^{o},\ a_2=a_2^{o},\ \ldots,\ a_r=a_r^{o},\ a_{r+1}=a_{r+1}^{o}.$$

Consider the column vector $\underline{a}^{o}$ of k rows defined as follows:

$$\underline{a}^{o} = (a_1^{o},\ a_2^{o},\ \ldots,\ a_r^{o},\ a_{r+1}^{o},\ 0,\ 0,\ \ldots,\ 0)$$

Clearly

$$\underline{u}_i\underline{a}^o = 0 \qquad \text{for } i = 1,2,\ldots,r \qquad (5)$$

$$\underline{u}_m\underline{a}^o = 1$$

In order to prove that $\underline{a}^o$ is a solution of the system of equations (2), it remains to show that (5) holds also for i=r+1, r+2, ..., k-1 as well. However, it follows from (3) that

$$\underline{u}_i\underline{a}^o = \lambda_{i1}\underline{u}_1\underline{a}^o + \lambda_{i2}\underline{u}_2\underline{a}^o + \ldots + \lambda_{ir}\underline{u}_r\underline{a}^o = 0 \quad \text{for } i=r+1,r+2,\ldots,n$$

On the other hand, as a result of condition 3 of theorem 2, there exist scalars such that

$$\underline{u}_i = \underline{u}_m + \lambda_{i1}\underline{u}_1 + \lambda_{i2}\underline{u}_2 + \ldots + \lambda_{in}\underline{u}_n \quad \text{for } i=n+1,\ldots,m$$

Taking the scalar product with $\underline{a}^o$ it now follows immediately that

$$\underline{u}_i\,\underline{a}^o = 1 \qquad \text{for } i = n+1,\ldots,m.$$

This completes the proof of the first part of theorem 2 concerning the existence of a solution of (2) whenever the conditions of theorem 2 are satisfied.

Conversely, suppose that (2) has a solution, $\underline{a}^o$, but in the sequence of testing for residual disclosure no earlier solution was found (i.e. no union of fewer than m-r elementary intersection sets would be a residual disclosure).

Let the rank of N be r. We will first show that in this case the rank of M is greater than r (in which case it will have to be r+1).

N has r linearly independent rows. Suppose, without loss of generality, that these are the first r rows. If the rank of M was not greater than r then it would have to be equal to r and so all rows of M would be linear combinations of the first r rows. In particular, there would exist scalars λ_i such that

$$\underline{u}_m = \lambda_1\underline{u}_1 + \lambda_2\underline{u}_2 + \ldots + \lambda_r\underline{u}_r \quad ; \quad r \leq n<m \qquad (6)$$

Since $\underline{a}^o$ is a solution of (2), we have

$$\underline{u}_i\underline{a}^o = 0 \qquad \text{for } i=1,2,\ldots,r,r+1,\ldots,n \tag{7}$$

$$\underline{u}_m\underline{a}^o = 1 \tag{8}$$

Taking the scalar product of (6) with $\underline{a}^o$, we obtain

$$\underline{u}_m\underline{a}^o = \lambda_1\underline{u}_1\underline{a}^o - \lambda_2\underline{u}_2\underline{a}^o + \ldots + \lambda_r\underline{u}_r\underline{a}^o \tag{9}$$

According to (8) the left hand side of (9) is equal to 1, while according to (7) the right hand side of (9) is equal to zero.

Clearly, the assumption that M has the same rank as N leads to a contradiction. Hence M must have a greater rank.

If M has a rank equal to r+1 then it follows, as in the proof of the first half of the theorem that there are r+1 rows (r rows of N plus one of the last m-n rows of M) such that all other rows of M are linear combinations of these. Without loss of generality, assume that these are the first r rows of M plus its last row. Then there are scalars λ_{ij} such that

$$\underline{u}_i = \lambda_{i1}\underline{u}_1 + \lambda_{i2}\underline{u}_2 + \ldots + \lambda_{ir}\underline{u}_r + \lambda_{im}\underline{u}_m \qquad i=n+1,\ldots,m$$

Taking the scalar product of $\underline{u}_i$ with $\underline{a}^o$, we get

$$1 = \lambda_{i1}\underline{u}_1\underline{a}^o + \lambda_{i2}\underline{u}_2\underline{a}^o + \ldots + \lambda_{ir}\underline{u}_r\underline{a}^o + \lambda_{im}\underline{u}_m\underline{a}^o = \lambda_{im}$$

This completes the proof of condition 3 of theorem 2.

If M has a rank greater than r+1, say r+t (t>1) then we can increase the rank of N by adding to it a suitably chosen row from among the last m-r. Continue adding to N rows from among the last m-r until we obtain a matrix whose rank is r+t-1; add to N also all rows from among the remaining last m-r rows of M which can be added without increasing the rank of N. The resulting matrix N' would satisfy all conditions of theorem 2, so the corresponding set of equations has a solution. It is easy to see that in this case we should have encountered residual disclosure corresponding to fewer than m-n elementary intersection sets and hence the process of testing would have stopped before. This completes the proof of theorem 2.

In order to prove theorem 3.b., observe that if every row of M is linearly dependent on the other rows then the omission of no single row can reduce the rank of M. Hence the rank of N (obtained from M by the deletion of a single row) is equal to the rank of M, hence (2) has no solution and there is no residual disclosure. Conversely, if a row of M, say the i-th, is linearly independent of the others, then the matrix N obtained by the deletion of the i-th row of M has smaller rank than M, so (2) has a solution. If in addition the aggregate corresponding to B_i is a direct disclosure, then it follows that the publication of the aggregates A_1, A_2, ..., A_k constitutes a residual disclosure.

REFERENCES

[1] Decennial Census Review Committee: "The Decennial Census". Report to the U.S. Secretary of Commerce, July, 1971.

[2] "Privacy and Confidentiality in the Federal Statistical System". United States, panel I Report, 1971, p10.

[3] The Globe and Mail, December 16, 1971, p1.

[4] Interview with W.E. Duffet, Chief Statistician of Canada and I.P. Fellegi, Director General, Methodology and Systems Branch, Statistics Canada.

[5] Canning, R.G.: "Data Security in the Corporate Data Base". EDP Analyzer, May, 1970, p12.

[6] Fellegi, I.P.: "On the Question of Statistical Confidentiality". Proceedings of the Social Statistics Section, American Statistical Association, 1970, p7.

[7] See reference [6], p10.

[8] Fellegi, I.P. and Vander Noot, T.J.: "Statistical Data Banks in the Canadian Government and Their Use". Statistics Canada, unpublished paper, 1971, p10.

[9] See reference [6], p12.

[10] See reference [4].

[11] Boruch, R.F.: "ACE Research and the Confidentiality of Data". Proceedings of the Social Statistics Section, American Statistical Association, 1969, p415.

[12] See reference [11], p413.

[13] Bachi, R. and Baron, R.: "Confidentiality Problems Related to Data-Banks". IAG Quarterly Journal, Vol. 2, No. 3, p44.

[14] "PERSONNENKENNZEICHNEN", a booklet published by the Federal Department of the Interior, Bonn, Federal Republic of Germany, June, 1971.

[15] Intergovernmental Council for ADP. Document No. GC-75, 1971.

[16] See Reference [14].

[17] See Reference [15].

[18] The Economist, September 6, 1969, p53.

[19] "A National Survey of the Public's Attitudes Towards Computers". A joint project of the American Federation of Information Processing Societies and Time Magazine, 1971.

[20] Canadian Chartered Accountant, July, 1970, p11.

[21] Canadian Datasystems, July, 1971, p53.

[22] Balmer, D., Chairman, Standards Committee on the Representation of Data Elements, Canadian Standards Association. Canadian Datasystems, October, 1971, p11.

[23] Miller, A.R.: The Assault on Privacy. University of Michigan Press, 1971, p22.

[24] Report on Protection of Privacy in Ontario, Ontario Law Reform Commission, 1968, p79.

[25] Warner, M. and Stone, M.: The Data Bank Society. George Allen & Unwin Ltd., 1970, p71.

[26] Martin, J. and Norman, A.R.D.: The Computerized Society. Prentice-Hall, 1970, p73.

[27] See reference [4].

[28] Telephone interview with Prof. C.M. Lanphier, Director, Survey Research Centre, Institute for Behavioural Research, York University.

[29] McPhail, T.: "Social Science Research and the Rights of Human Subjects". Report to the Privacy and Computers Task Force, p7.

[30] Hellman, J.J.: "Privacy and Information Systems: An Argument and an Implementation". The Rand Corporation, May, 1970, p31.

[31] See reference [1] p6.

[32] Ottawa Citizen, October 6, 1971, p3.

STUDIES COMMISSIONED BY THE TASK FORCE

The Nature of Privacy - D.N. Weisstub and C.C. Gotlieb.

Personal Records: Procedures, Practices, and Problems - J.M. Carroll and J. Baudot, Carol Kirsh, J.I. Williams.

Electronic Banking Systems and Their Effects on Privacy - H.S. Gellman.

Technological Review of Computer/Communications.[1]

Systems Capacity for Data Security - C.C. Gotlieb and J.N.P. Hume.

Statistical Data Banks and Their Effects on Privacy - H.S. Gellman.

Legal Protection of Privacy - J.S. Williams.

Vie Privée et Ordinateur Dans le Droit de la Province du Québec - J. Boucher.

Regulation of Federal Data Banks - K. Katz.

Regulatory Models - J.M. Sharp.

Ordinateur et Vie Privée: Techniques et Contrôle - C. Fabien.

The Theory and Practice of Self-Regulation - S.J. Usprich.

Privacy, Computer Data Banks, Communications and the Constitution - F.J.E. Jordan.

International Factors - C. Dalfen.

[1] A joint Study by the Privacy and Computers Task Force and the Canadian Computer/Communications Task Force, to be published by the latter.