

**STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE  
EUROPEAN COMMUNITIES**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROSTAT**

**Joint ECE/Eurostat Work Session on  
Statistical Data Confidentiality**

(Skopje, The former Yugoslav Republic of Macedonia,  
14-16 March 2001)

Working Paper No. 9  
English only

Topic I: Application of statistical disclosure control methodology and software in business statistics and social and demographic statistics

## **A DECISION-THEORETIC APPROACH TO DATA DISCLOSURE PROBLEMS**

### **Contributed paper**

Submitted by the Universitat de Valencia (Spain) and Carnegie Mellon University (United States)<sup>1</sup>

**Abstract:** A major concern of statistical agencies is to protect the confidential information contained in data sets that have to be released. The problem is not trivial since the agency has to find an optimal balance between the needs of the users and the needs of the data providers. Ideally the statistical agency should provide maximum information to the users, preserving the privacy of the individual entities represented in the data. The sub-field of statistic concern with such a problem is usually referred to as *statistical data protection*. In this paper I present a decision-theoretic approach to data disclosure problems. The approach is innovative because (i) it offers a theoretical framework to develop optimality criteria for the choice of the best form of data release, (ii) it recognizes the different perspectives of the statistical agency and of the users of the data in assessing the extent of disclosure and the quality of the users inference (data utility) associated with different forms of data release. This leads to new measures of the extent of disclosure and data utility and a new interpretation of the notion of *inferential disclosure* currently used in statistical confidentiality.

### **I. INTRODUCTION**

1. As a part of their activities, most European statistical agencies release data sets containing information on individual entities subject to pledges of confidentiality. Nowadays confidentiality is a major legal concern for all statistical agencies as a consequence of the European Union's Data Protection Directive, as well as the laws governing privacy and confidentiality of statistical data in individual countries. In the last ten years, the amount of statistical data collected has increased enormously and new statistical algorithms and expanding power of computers have increased the danger of disclosure of confidential information. The more statistical data are collected, in fact, the easier it becomes to match a data set of interest with other data sets containing information that allow to identify the records in the data set of interest. Confidentiality protection is not only an ethical problem, it is also crucial for the quality and availability of statistical data. The willingness of individual entities to participate in statistical data collection depends, in fact, on the assurances that the data will be released in a form that won't harm them. On the other hand, statistical information has become a key element for the actions that both private and public decision makers have to take, with a corresponding increase in the demand of release of statistical data. Government agencies use statistical data to decide the allocation of funds and to monitor social programmes; policy analysts use statistical data to inform social decisions; researchers use statistical data to test their theories and to achieve a deeper understanding of the phenomena under study. Availability of statistical data is therefore of great importance to improve the quality of the decisions and to speed up the research in many different fields of great interest for modern society.

---

<sup>1</sup> Prepared by Mario Trottini.

2. Ideally a statistical agency should provide maximum information to the users preserving the privacy of the individual entities represented in the data set. The sub-field of statistics concerned with such a problem is usually referred to as *statistical data protection*, *statistical confidentiality* or *statistical disclosure limitation*.

3. In a typical disclosure scenario an intruder attempts to infer a target value based on the information contained in a data set that has to be released. Statistical disclosure limitation then involves:

- (i) the assessment of the *extent of disclosure*, i.e. the extent to which the intruder is able to infer the target value from the released data set;
- (ii) classification of the data set as “safe” if the extent of disclosure is small or “unsafe” if the extent of disclosure is high;
- (iii) modification of the original data set using different disclosure limitation techniques in order to obtain a safer data set;
- (iv) assessment of the impact of the disclosure limitation techniques used on the quality of the data and therefore on the inference of typical users;
- (v) choice of the best form of data release based on the trade-off of gains versus risk.

4. The major efforts, in statistical confidentiality to date have been devoted to the assessment of the extent of disclosure (see for example Duncan and Lambert, 1986), and to the development of disclosure limitation techniques, that determine the content of data sets or files to be released. If the data to be released have an associated risk of disclosure that is deemed to be too high (with respect to some threshold value), then statistical agencies apply disclosure limitation techniques, such as deleting identifiers (name, social security number, address), dropping sensitive variables, releasing only a subset of observations, etc. (Duncan and Pearson, 1991). Very little has been done so far, however, to measure the impact of these techniques on statistical analyses and to assess the trade-off of gains versus risk (Fienberg, 2000). The use of disclosure limitation techniques can seriously alterate relations among study variables and result in analyses that are incomplete or misleading (Kamlet et al., 1985). In the most extreme case if we suppress all values in a data set, we obtain a new data set which is certainly safe but contains no information. Further, there have been few efforts to develop formal optimality criteria for the choice of data release. In practice the choice is currently made based on heuristic considerations. The dangers of this lack of theoretical roots of the current approaches is well established (see for example recommendation 6.2 in the Panel on Confidentiality and Data Access (Duncan et al., 1993)).

5. This paper presents a decisional-theoretic approach to data disclosure problems that takes into account both the extent of disclosure and the quality of the users' inference (data utility) associated with different forms of data release. In our disclosure scenario it is assumed that the set of users can be partitioned in two groups. Those who want to use the released data to perform statistical studies or for research purposes, and those who want to use the released data to disclose confidential information about the providers of the data. I will refer to the first group as *society* and to the second group as the *intruder*. More formally, it is assumed that society is interested in a unknown quantity  $\Theta_{SOC}$  while the intruder is interested in an unknown quantity  $\Theta_{INT}$  and they try to infer their target values,  $\Theta_{SOC}$  and  $\Theta_{INT}$ , based on the information contained in the released data. Ideally the statistical agency should provide maximum information about  $\Theta_{SOC}$  and minimum information about  $\Theta_{INT}$ . In general this is not feasible since there is a trade-off between the two types of information and it is not possible to increase one without also increasing the other.

6. The approach presented in the paper generalizes and extends previous work of Duncan and Lambert (1986) and Duncan and Keller-McNulty (1999). The approach is innovative in two respects: (i) it offers a theoretical framework to measure the data utility and to define optimality criteria that go beyond the few ad-hoc procedures that have been proposed so far; (ii) it recognizes the different perspectives of the statistical agency and of the users of the data in assessing the extent of disclosure and the data utility. This leads to a new interpretation of the notion of *inferential disclosure* currently used in research on statistical confidentiality, and a more complex and, I believe, more appropriate formalization, of the problem of statistical confidentiality.

7. In Section II, I review in some detail the work of Duncan and Lambert (1986) and Duncan and Keller-McNulty (1999) that inspired the main ideas of this work. Section III sets up the basic assumptions and notation. In section IV, I introduce a new definition of disclosure and data utility and describe a general framework for the representation of the trade-off of gains versus risk. Also in section IV an optimality criterion for the choice of the best form of data release is presented. Section V contains concluding remarks and outlines plans for future work.

## II. CURRENT MEASURES OF THE EXTENT OF DISCLOSURE AND DATA UTILITY

8. Among the attempts to measure the extent of disclosure the most successful, so far, seems to be the approach proposed by Duncan and Lambert (1986). In their work the authors provide a decision-theoretic framework to measure the extent of disclosure that includes as special cases, for suitable choices of the loss function, many ad-hoc criteria proposed for the problem. The approach is as simple as powerful. The different measures of disclosure are obtained applying an uncertainty function to the prior and posterior distributions of the intruder. An uncertainty function is a non-negative measurable mapping from the space of all possible distributions to the set of non-negative real number. Each distribution is associated with a non-negative number. The bigger is the number the bigger is the uncertainty about the value of the random variable with that distribution. In particular the authors define three classes of disclosure measures: (intruder's) *knowledge*, (intruder's) *knowledge gain*, and (intruder's) *relative knowledge gain*. The posterior knowledge measure of disclosure,  $U(\text{posterior})$ , is obtained applying the uncertainty function  $U(\bullet)$  to the intruder's posterior distribution for  $\Theta_{INT}$ . This is the right quantity to consider if the final goal is to protect the intruder's knowledge about the target. However, if the intruder already has precise information about the target value before the release of the data, the knowledge gain, i.e. the difference  $U(\text{prior})-U(\text{posterior})$ , or the relative knowledge gain,  $(U(\text{prior})-U(\text{posterior}))/U(\text{prior})$ , might be more appropriate measures of disclosure. The intruder's loss function is assumed to be non-negative and bounded and the uncertainty function is defined as the expected loss associated with the intruder's optimal action. DeGroot, 1962, showed that this is a very general class of uncertainty functions. Many measures of the extent of disclosure currently used by statistical agencies can be obtained as a special case of this general framework for suitable choices of the loss function.

9. While considerable research has been devoted to the assessment of the extent of disclosure, very little has been done to measure the impact of statistical disclosure techniques on the statistical analyses and to assess the trade-off of gains versus risk. Kamlet et al. (1985) provide examples of bias and inconsistency introduced by grouping and adding noise, and show how alternative methods can be used to properly analyzed the masked data. The Panel on Confidentiality and Data Access, note 6, page 178 (Duncan et al., 1993), illustrates situations in which adding noise has no effect in important kinds of analyses. Other works by Sanz and Domingo-Ferrer (1998) and Baeyens and Defays (1998) discuss measures of information loss for microaggregation methods. De Wall and Willesborg (1998) and Hurkens and Tiourine (1998) consider measures of information loss for local suppression and global recoding. Although very interesting this prior research offers solutions that are problem specific and does not provide a general framework to measure the data utility and to assess the trade-off of gain versus risk. The work of Duncan and Keller-McNulty represents a first tentative to build such a framework. In their method the extent of disclosure is measured as in Duncan and Lambert (1986) and the data utility is measured in terms of the MSE for the estimator of society's target. The optimal form of data release is the one that maximizes the data utility among those with extent of disclosure below a fixed threshold. The emphasis of the work is on the representation of the trade-off of gain versus risk rather than on the development of an appropriate measure of data utility. The choice of the MSE is made only for convenience and is not intended by the authors as a general way to measure the data utility. The author has been working on an implementation of the approach described by Duncan and Keller McNulty in which the measure of data utility is obtained as extension of the measures of disclosure discussed in Duncan and Lambert (1986), instead of in terms of MSE. In particular applying an uncertainty function to the prior and posterior distributions of the society I defined three new measures of data utility, *society's knowledge*, *society's knowledge gain*, *society's relative knowledge gain* whose definition and interpretation mirror the one given for the corresponding measures of disclosure. In the framework described by Duncan and Lambert these seems to be a much natural way to measure the data utility than the MSE. Despite the progress and the advances in statistical confidentiality in last the 30 years many questions are still open. In particular the role of the statistical agency in assessing the risk of disclosure and the data utility has

undergone very little exploration. The existing measures of disclosure and data utility are obtained trying to reproduce what would be the intruder's and society's behavior when a data set is released. However they do not take into account the agency knowledge of the intruder's and society's targets. The author strongly believes that the agency's perspective should be also a component of the measure of the extent of disclosure and data utility. Consider the following example.

### Example 1

10. Suppose that two data sets  $D_1$  and  $D_2$ , resulting from two different disclosure limitation techniques, are considered for release. Suppose that the intruder's posterior distributions for  $\Theta_{INT}$  given  $D_1$  and  $D_2$  are  $N(-5,1)$  and  $N(5,1)$  respectively while the society's posterior distributions for  $\Theta_{SOC}$  given  $D_1$  and  $D_2$  are  $N(-7,2)$  and  $N(7,2)$  (here  $N(m,v)$  denotes a normal distribution with mean  $m$  and variance  $v$ ). Suppose also that the agency's posterior distributions for  $\Theta_{INT}$  and  $\Theta_{SOC}$  given the original data are  $N(4.9,0.001)$  and  $N(-6.9,0.001)$  respectively and that intruder, society and agency use a quadratic loss function. Under quadratic error loss the optimal action is the posterior mean and the uncertainty (i.e. the expected loss associated with the optimal action) is the posterior variance. Thus the intruder's optimal estimates of  $\Theta_{INT}$  when  $D_1$  and  $D_2$  are released are -5 and 5 respectively while the intruder's uncertainty in both cases is 1. Similarly the society's optimal estimates of  $\Theta_{SOC}$  when  $D_1$  and  $D_2$  are released are -7 and 7 and the society's uncertainty in both cases is 2. Based on the current measures of the extent of disclosure and their natural extension to data utility the agency should flip a coin to decide which data set is worth to release. Either the agency releases  $D_1$  or the agency releases  $D_2$ , the intruder's uncertainty is 1 and the society's uncertainty is 2 and therefore the intruder's and society's posterior knowledge, the knowledge gain and the relative knowledge gain of  $D_1$  and  $D_2$  are the same. Using the representation of the trade-off of gain versus risk proposed by Duncan and Keller-McNulty (1999)  $D_1$  and  $D_2$  are perfectly equivalent. However the agency posterior distribution for  $\Theta_{INT}$  and  $\Theta_{SOC}$  are  $N(4.9,0.001)$  and  $N(-6.9,0.001)$ . Thus the agency is very confident that the true value of  $\Theta_{INT}$  is approximately 4.9 (the 95% posterior credibility interval for  $\Theta_{INT}$  is [4.84,4.96]) and the true value of  $\Theta_{SOC}$  is -6.9 (the 95% posterior credibility interval for  $\Theta_{SOC}$  is [-6.96,-6.84]). If  $D_2$  is released the intruder's optimal estimate of  $\Theta_{INT}$  is 5, very close to what the agency's believes to be the true value of  $\Theta_{INT}$  while the society's optimal estimate of  $\Theta_{SOC}$ , 7, is very poor from the agency's point of view. If  $D_1$  is released, instead, the intruder's optimal estimate of  $\Theta_{INT}$ , -5, is very inaccurate from the agency's point of view while the society's optimal estimate of  $\Theta_{SOC}$ , -7, is very close to what the agency believes to be the true value of  $\Theta_{SOC}$ . Thus if we take into account not only the intruder's and society's perspectives, but also the agency's knowledge of the intruder's and the society's targets  $\Theta_{INT}$  and  $\Theta_{SOC}$ , the two forms of data release,  $D_1$  and  $D_2$ , are not longer equivalent and  $D_1$  should be intuitively released, since the intruder's and the society's uncertainties given  $D_1$  or given  $D_2$  are the same but, from the agency's perspective, if  $D_2$  is released the intruder's inference is very precise and the society's inference is very poor while if  $D_1$  is released the intruder's inference is very imprecise and the society's optimal estimate is very close to what the agency believes to be the true value of  $\Theta_{SOC}$ . Based on the approach proposed by Duncan and Keller-McNulty (1999) it is not possible to distinguish between the release of  $D_1$  and the  $D_2$ . In section 4 we present new measures of the extent of disclosure and data utility that allows this distinction. Based on this new measures of extent of disclosure and data utility the release of  $D_1$ , according with the intuition, is preferred to the release of  $D_2$ .

11. In the next section we introduce the basic notation that will be used throughout the paper and we state the assumptions characterizing the new measures of the extent of disclosure and data utility that we propose in section IV.

### III. NOTATION AND GENERAL ASSUMPTIONS

12. As a result of its activities, a statistical office produces a data set  $D_0$ . In order to take into account confidentiality issues and the needs of the users of the data, different forms of data release are considered. The statistical office can release the original data  $D_0$  decide not to release any data, or decide to apply *disclosure limitation techniques* and release modified data or synthetic data. We denote with  $D_R$  the released data set and we assume that  $D_R$  belongs to a set  $\mathbf{D}$  of alternative forms of data release. Ideally the statistical agency should release maximum information about  $\Theta_{SOC}$  and minimum information about  $\Theta_{INT}$ .

13. The definition of an optimality criterion requires some assumptions about the behavior of the intruder, society, and the statistical agency, how they formalize their prior information about the target values, how they update this prior information, how they use the released data set to make inference etc. In order for the problem to be meaningful, it is assumed that both  $\Theta_{INT}$  and  $\Theta_{SOC}$  are somehow related to the original data  $D_0$ . In particular, it is assumed that, prior to observing  $D_0$ , the statistical agency believes that  $D_0$  is a realization of a random variable  $V_A$  whose distribution  $P_A$  belongs to a parametric family  $\mathbf{PA}$  with parameter  $\Psi_A$  and parameter space  $\Omega_A$ . Similarly it is assumed that the intruder and society, prior to observing  $D_0$ , believe that  $D_0$  is a realization of random variables  $V_{INT}$  and  $V_{SOC}$  whose distributions  $P_{INT}$  and  $P_{SOC}$  belong to parametric families  $\mathbf{PINT}$  and  $\mathbf{PSOC}$  with parameters  $\Psi_{INT}$  and  $\Psi_{SOC}$  and parameter spaces  $\Omega_{INT}$  and  $\Omega_{SOC}$  respectively. The conditional distributions of  $V_A$ ,  $V_{INT}$ , and  $V_{SOC}$  given  $\Theta_{INT}$ , and the conditional distributions of  $V_A$ ,  $V_{INT}$ , and  $V_{SOC}$  given  $\Theta_{SOC}$ , formalize how the data set  $D_0$  is related to the intruder's and society's targets,  $\Theta_{INT}$  and  $\Theta_{SOC}$ . It is also assume that the intruder's and society's prior uncertainties about  $\Psi_{INT}$ ,  $\Theta_{INT}$ , and  $\Psi_{SOC}$ ,  $\Theta_{SOC}$ , can be adequately expressed by probability distributions,  $\pi_{\Psi_{INT}}(\bullet)$ ,  $\pi_{\Theta_{INT}}^{(I)}(\bullet)$ ,  $p_{\Psi_{SOC}}(\bullet)$ ,  $p_{\Theta_{SOC}}^{(A)}(\bullet)$ , that I will refer to as *intruder's and society's prior distributions* for  $\Psi_{INT}$ ,  $\Theta_{INT}$ , and  $\Psi_{SOC}$ ,  $\Theta_{SOC}$  respectively. Similarly it is assume that the agency's prior uncertainty about  $\Psi_A$ ,  $\Theta_{INT}$ , and  $\Theta_{SOC}$  can be adequately expressed by probability distributions,  $\pi_{\Psi_A}(\bullet)$ ,  $\pi_{\Theta_{INT}}^{(A)}(\bullet)$ ,  $p_{\Theta_{SOC}}^{(A)}(\bullet)$  that I will refer to as *agency's prior distributions* for  $\Psi_A$ ,  $\Theta_{INT}$ , and  $\Theta_{SOC}$  respectively. The conditional distributions of  $V_{INT}$  given  $\Psi_{INT}$ , and  $V_{SOC}$  given  $\Psi_{SOC}$ , the prior distributions for  $\Psi_{INT}$  and  $\Psi_{SOC}$ , and the particular disclosure limitation technique used, induce, for each form of data release,  $D_R$ , conditional distributions of  $D_R$  given  $\Theta_{INT}$  and  $D_R$  given  $\Theta_{SOC}$  that formalize how the released data set  $D_R$  is related to the intruder's and society's targets  $\Theta_{INT}$  and  $\Theta_{SOC}$ . We denote by  $\pi_{\Theta_{INT}}^{(I)}(\bullet|D_R)$  the conditional distribution of  $\Theta_{INT}$  given  $D_R$  and by  $\pi_{\Theta_{SOC}}^{(S)}(\bullet|D_R)$  the conditional distribution of  $\Theta_{SOC}$  given  $D_R$ . These are the *intruder's and society's posterior distributions* for  $\Theta_{INT}$  and  $\Theta_{SOC}$  given  $D_R$  and they express the complete intruder's and society's uncertainties about their targets, after the data  $D_R$  have been released. Similarly, we denote by  $\pi_{\Theta_{INT}}^{(A)}(\bullet|D_0)$  and  $\pi_{\Theta_{SOC}}^{(A)}(\bullet|D_0)$  the conditional distributions of  $\Theta_{INT}$  and  $\Theta_{SOC}$  given  $D_0$ . These are the *agency's posterior distributions* for  $\Theta_{INT}$  and  $\Theta_{SOC}$  given  $D_0$  and they express the complete agency's uncertainty about the target values  $\Theta_{INT}$  and  $\Theta_{SOC}$ , after the data  $D_0$  have been observed. Note that the intruder's and society's posterior distributions are usually different from the agency's posterior distributions. This is not just because the agency's priors for  $\Theta_{INT}$  and  $\Theta_{SOC}$ , might be different from the intruder's and society's priors, or because the agency's model might be different from the intruder's and society's models but also because the statistical agency possesses the original data  $D_0$  and therefore updates its uncertainty about  $\Theta_{INT}$  and  $\Theta_{SOC}$  using the original data  $D_0$  instead of the released data  $D_R$ . We also assume the following:

**ASSUMPTION 1:** The statistical agency knows the intruder's and society's loss functions, their prior distributions for the targets values  $\Theta_{INT}$ ,  $\Theta_{SOC}$  as well as their uncertainties about the model generating the original data  $D_0$ .

14. Assumption 1 is not very realistic; however it can be easily relaxed to fit more realistic scenarios. For example the statistical office can use classes of prior distributions and classes of loss functions to describe the intruder's and society's prior uncertainties about their targets, and the loss that the intruder and society are willing to pay for a generic estimate of their target values. In a similar way classes of distributions can be used to describe the intruder's and society's uncertainties about the mechanism producing the data  $D_0$ .

**ASSUMPTION 2:** The statistical office release complete information about the mechanism that produces the released data set.

15. In real life assumption 2 is usually not satisfied, but I believe it should be. In most of the cases statistical offices do not release complete information about the disclosure limitation technique used. For example the concentration rule for cell suppression is often not revealed to data users (Duncan et al., 1993). In mine opinion, however, the statistical office should release as much information as possible about the mechanism generating the released data set. The more information that is available the easier it is to predict the behavior of the intruder and society and therefore to make correct inferences about the disclosure risk and the data utility associated with the released data.

**ASSUMPTION 3:** Both the intruder and society act rationally, according to the expected loss principle, i.e. in estimating their targets they try to minimize the posterior risk.

16. The approach is normative. We describe what the intruder and society should do rather than what they actually would do. From the statistical agency's point of view, a descriptive approach might seem more appropriate. However a descriptive approach is much more difficult to implement and when assumption 2 is satisfied there should not be a big difference between the two approaches.

**ASSUMPTION 4:** The intruder takes actions only when his/her uncertainty about the target  $\Theta_{INT}$  is below a fixed threshold  $k_{INT}$ .

17. Assumption 4 implicitly assumes that the intruder has to pay a penalty when he/she claims that confidential information has been disclosed, but actually no disclosure has taken place. This assumption is quite realistic in many problems although not in general.

Representing Disclosure Risk and Data Utility

The choice of the best form of data release requires a notion of disclosure risk and data utility. In a broad sense the disclosure risk associated with the release of a data set  $D_R$  is a measure of the extent to which the release of  $D_R$  makes it possible for the intruder to harm. It can be that the intruder's goal is to obtain information about particular individuals or to discredit the data provider or just show his/her own cleverness. In most of the cases these goals co-exist and it safe for the statistical agency to act "as if" all these goals actually co-exist. As a result we make the following assumptions:

**ASSUMPTION 5:** The intruder can create harm in two different ways:

**Disclosure harm:** the intruder discloses confidential information about the providers of the data, i.e. the intruder's inference is correct;

**Discredit harm:** the intruder discredits the statistical agency or the data providers, claiming that confidential information has been disclosed.

18. Note that while the disclosure harm only takes place when the intruder's inference about  $\Theta_{INT}$  is correct, (for example when the intruder correctly identifies the owner of some record in a released microdata set), the discredit harm can occur even when the intruder's inference is completely incorrect but the intruder believes that his/her inference is very precise and acts in consequence. Also note that while under assumptions 1-5 the statistical agency knows the probability of discredit harm (since this depends on the intruder's uncertainty), the agency, in general can only estimate the disclosure harm, since, in general, it does not have perfect information about the true value of the intruder's target  $\Theta_{INT}$ . The two types of harm can occur in different ways. We can have only disclosure harm or only discredit harm, or both. We need to distinguish between these three situations. We introduce the following definitions (in what follows  $U_{\Theta_{INT}}^{(I)}(\bullet)$ ,  $U_{\Theta_{INT}}^{(A)}(\bullet)$ ,  $U_{\Theta_{SOC}}^{(S)}(\bullet)$  are uncertainty functions as in Duncan and Lambert, 1986).

**DEFINITION 1:** Let  $U_{\Theta_{INT}}^{(I)}(p^{(I)})$  denote the intruder's uncertainty about  $\Theta_{INT}$  when his/her distribution over  $\Theta_{INT}$  is  $p^{(I)}(\bullet)$ . We define the *risk of discredit harm* when the intruder distribution over  $\Theta_{INT}$  is  $p^{(I)}(\bullet)$  as:

$$\text{RDH}(p^{(I)}) = -U_{\Theta_{INT}}^{(I)}(p^{(I)}).$$

19. Definition 1 somehow rephrases and specializes assumption 4. The risk of discredit harm is a decreasing function of the intruder's posterior uncertainty about  $\Theta_{INT}$ . Here I used the function "minus the uncertainty", but other choices, of course, are possible. Note that RDH is minus the posterior knowledge measure of disclosure proposed by Duncan and Lambert (1986). Note also that  $\text{RDH} \leq 0$  and  $\text{RDH} = 0$  if and only if the intruder's has no uncertainty about the target  $\Theta_{INT}$ .

**DEFINITION 2:** If for a threshold  $t_{INT}$ ,

$$(i) \text{RDH}(\pi_{\Theta_{INT}}^{(I)}(\bullet)) < t_{INT};$$

$$(ii) \text{RDH}(\pi_{\Theta_{INT}}^{(I)}(\bullet | D_R)) > t_{INT};$$

then we say that *disclosure from the intruder's point of view* has taken place.

**DEFINITION 3:** Let  $G_A$  and  $G_{INT}$  be two generic probability distributions, and let  $D^{(A)}(G_A, G_{INT})$  be an arbitrary measure of divergence between  $G_A$  and  $G_{INT}$ .  $D^{(A)}(G_A, G_{INT})$  is a measure of how well  $G_{INT}$  approximates  $G_A$ . Also let  $U_{\Theta_{INT}}^{(A)}(p^{(A)})$  be the agency's uncertainty about  $\Theta_{INT}$  when its distribution over  $\Theta_{INT}$  is  $p^{(A)}(\bullet)$ . We define the *intruder's estimated knowledge* (IEK) from the agency's point of view when the intruder's distribution over  $\Theta_{INT}$  is  $p^{(I)}(\bullet)$  as:

$$\text{IEK}(p^{(I)}) = -[D^{(A)}(\pi_{\Theta_{INT}}^{(A)}(\bullet|D_0), p^{(I)}) + U_{\Theta_{INT}}^{(A)}(\pi_{\Theta_{INT}}^{(A)}(\bullet|D_0))].$$

In particular we define *the intruder's prior estimated knowledge*  $\text{IEK}(\pi_{\Theta_{INT}}^{(I)}(\bullet))$ , and the *intruder's posterior estimated knowledge*,  $\text{IEK}(\pi_{\Theta_{INT}}^{(I)}(\bullet|D_R))$ .

20. Definition 3, says that, from the agency's point of view, the intruder's knowledge about  $\Theta_{INT}$  when his/her distribution over  $\Theta_{INT}$  is  $p^{(I)}$ , is extensive (i.e. the risk of disclosure harm is high) only if the agency's uncertainty about  $\Theta_{INT}$  based on the original data  $D_0$  is very small and the intruder's distribution over  $\Theta_{INT}$ ,  $p^{(I)}$ , does not differ too much from the agency's posterior distribution over  $\Theta_{INT}$  based on the original data  $D_0$ . Note that  $\text{IEK} \leq -U_{\Theta_{INT}}^{(A)}(\pi_{\Theta_{INT}}^{(A)}(\bullet|D_0)) \leq 0$  and  $\text{IEK} = 0$  if and only if the agency's has no uncertainty about the target  $\Theta_{INT}$  and the agency's and intruder's distributions over  $\Theta_{INT}$  are the same, as measured by the distance  $D^{(A)}$ .

**DEFINITION 4:** If for a threshold value  $t_A$ :

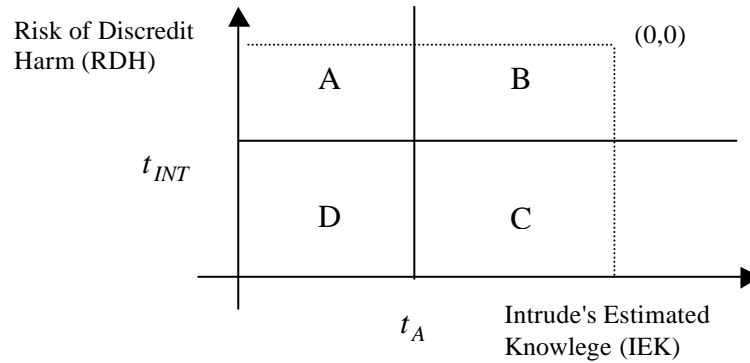
- (i) the intruder's prior estimated knowledge is less or equal than  $t_A$ ;
  - (ii) the intruder's posterior estimated knowledge is greater than  $t_A$ ;
- then we say that *disclosure from the agency's point of view* has taken place.

21. A reasonable measure of disclosure associated with the release of a data set  $D_R$  should take into account both the risk of disclosure harm and the risk of discredit harm. We propose the following definition of global risk.

**DEFINITION 5:** We define the *global risk* associated with the release of a data set  $D_R$  as the vector:

$$\text{G.RISK}(D_R) = [\text{IEK}(\pi_{\Theta_{INT}}^{(I)}(\bullet|D_R)), \text{RDH}(\pi_{\Theta_{INT}}^{(I)}(\bullet|D_R))]$$

22. Using definition 5, we can represent the risk associated with each form of data release  $D_R$  as a point  $p(D_R)$  in a Cartesian plane, with coordinates  $\text{G.RISK}(D_R)$ , as shown in figure 1.



**Figure 1:** *global risk*

23. Assuming that prior to the release of the data both the risk of discredit harm and the intruder's estimated knowledge are below the corresponding thresholds, points in A correspond to data sets for which the intruder's inference is very imprecise (from the agency's point of view), however, the intruder is very confident about his/her inference. In this case (from the agency's point of view) there is no violation of confidentiality but the intruder is very likely to take some actions and act "as if" confidentiality has been disclosed. We have discredit harm but not disclosure harm. Points in B correspond to data sets for which the intruder's inference is very precise both from the agency's and intruder's points of view. In this case (from the agency's point of view) the intruder is very likely to disclose confidential information and take actions

consequently. We have both disclosure harm and discredit harm. Points in C correspond to data sets for which the intruder's inference is correct (from the statistical office point of view), but the intruder is not very likely to take any action since his/her uncertainty about the target value is very high. For a point in C we have disclosure harm but not discredit harm. Similarly, points in D correspond to data sets for which there is not violation of confidentiality and the intruder is not likely to take any action, since his/her uncertainty about the target value is very high. Data sets in D are the safest, data sets that correspond to points in B are the most dangerous. In particular note that the point (0,0) corresponds to a data set whose release allows both the intruder and the agency to disclose the intruder's target with probability one. In a similar way we can define a global measure of data utility. In a broad sense the data utility associated with the release of a data set  $D_R$  is a measure of the extent to which the release of  $D_R$  makes it possible for society to make accurate inference about its target  $\Theta_{SOC}$ . Again we need to distinguish between the agency's and society's points of view. We introduce the following definitions.

**DEFINITION 6:** Denote with  $U_{\Theta_{SOC}}^{(S)}(p^{(S)})$  society's uncertainty about  $\Theta_{SOC}$  when its distribution over  $\Theta_{SOC}$  is  $p^{(S)}(\bullet)$ . We define the *data utility from society's point of view* (DUS) when society's distribution over  $\Theta_{SOC}$  is  $p^{(S)}(\bullet)$  as:

$$DUS(p^{(S)}) = -U_{\Theta_{SOC}}^{(S)}(p^{(S)}).$$

Note that  $DUS \leq 0$  and  $DUS=0$  if and only if society has no uncertainty about the target  $\Theta_{SOC}$ .

**DEFINITION 7:** Let  $D^{(A)}(\bullet, \bullet)$  be an arbitrary measure of divergence as in definition 3. Also let  $U_{\Theta_{SOC}}^{(A)}(p^{(A)})$  be the agency's uncertainty about  $\Theta_{SOC}$  when its distribution over  $\Theta_{SOC}$  is  $p^{(A)}(\bullet)$ . We define the *society's estimated knowledge* (SEK) from the agency's point of view, when society's distribution over  $\Theta_{SOC}$  is  $p^{(S)}(\bullet)$ , as:

$$SEK(p^{(S)}) = -[D^{(A)}(\pi_{\Theta_{SOC}}^{(A)}(\bullet | D_0), p^{(S)}) + U_{\Theta_{SOC}}^{(A)}(\pi_{\Theta_{SOC}}^{(A)}(\bullet | D_0))].$$

24. In particular we define the society's prior estimated knowledge,  $SEK(\pi_{\Theta_{SOC}}^{(S)}(\bullet))$ , and *the society's posterior estimated knowledge*,  $SEK(\pi_{\Theta_{SOC}}^{(S)}(\bullet | D_R))$ . Note that  $SEK \leq U_{\Theta_{SOC}}^{(A)}(\pi_{\Theta_{SOC}}^{(A)}(\bullet | D_0))$  and  $SEK=0$  if and only if the agency's has no uncertainty about  $\Theta_{SOC}$  and the agency's and society's distributions over  $\Theta_{SOC}$  are the same, as measured by the distance  $D^{(A)}$ .

**DEFINITION 8:** We define the *global data utility* associated with the release of a data set  $D_R$  as the vector:

$$G.UTILITY(D_R) = [SEK(\pi_{\Theta_{SOC}}^{(S)}(\bullet | D_R)), DUS(\pi_{\Theta_{SOC}}^{(S)}(\bullet | D_R))].$$

25. Note that the definitions of disclosure and data utility introduced here refer to the agency's, intruder's, and society's uncertainties about the target values  $\Theta_{INT}$ ,  $\Theta_{SOC}$ , and to the distance between the agency's the intruder's, and society's posterior distributions, but they do not refer to the information content of a data set or at least not in the usual sense. The information provided by data, in fact, is usually defined in terms of the changes produced in the probability distribution of interest (see for example Bernardo and Smith, 1994). We do not follow this approach because what really matters here is not how much information the released data set contains (measured in terms of distance between prior and posterior distribution of the target values) but rather to what extent its release allows the intruder and society to make accurate inferences about their targets. As a consequence, we can have that the released data set  $D_R$  contains a lot of information (prior and posterior distributions for  $\Theta_{INT}$  are very different) but still its release is safe.

26. Using the measures of global risk and global utility introduced above, we can define an optimality criterion for the choice of the best form of data release. We propose de following:

**Optimality Criterion:** Let  $D_{1A}$  be the subset of  $D$  containing all data set in  $D$  that have risk of discredit harm and intruder's estimated knowledge below the corresponding thresholds (i.e.  $D_{1A}$  consists of all data sets in  $D$  whose global risk belong to the region A in figure 1). If  $D_{1A}$  is empty then do not release any data set. If  $D_{1A}$  is not empty then release the data set in  $D_{1A}$  whose global utility minimize the Euclidean distance from the point (0,0).

27. The idea underlying the criterion is very simple. Among the *safe* data sets, i.e. those whose global risk is in the region A in figure 1, the optimal one is the one whose global utility is closest to the point of maximum global utility (0,0). If safe data set do not exist, no data set is released.

### Example 1 (continue)

28. Suppose that the agency uses as a measure of divergence between distribution the Kullback-Leibler divergence. Then the global risk of  $D_1$  is  $G.RISK(D_1) = (-51.96, -1)$  the global risk of  $D_2$  is  $G.RISK(D_2) = (-2.96, -1)$ , the global utility of  $D_1$  is  $G.UTILITY(D_1) = (-3.30, -2)$  and the global utility of  $D_2$  is  $G.UTILITY(D_2) = (-51.60, -2)$ . Figure 2 represents the trade-off of gain versus risk (threshold values  $t_{INT} = -0.8$  and  $t_A = -40$  for the risk of discredit harm and the intruder's estimated knowledge are used).

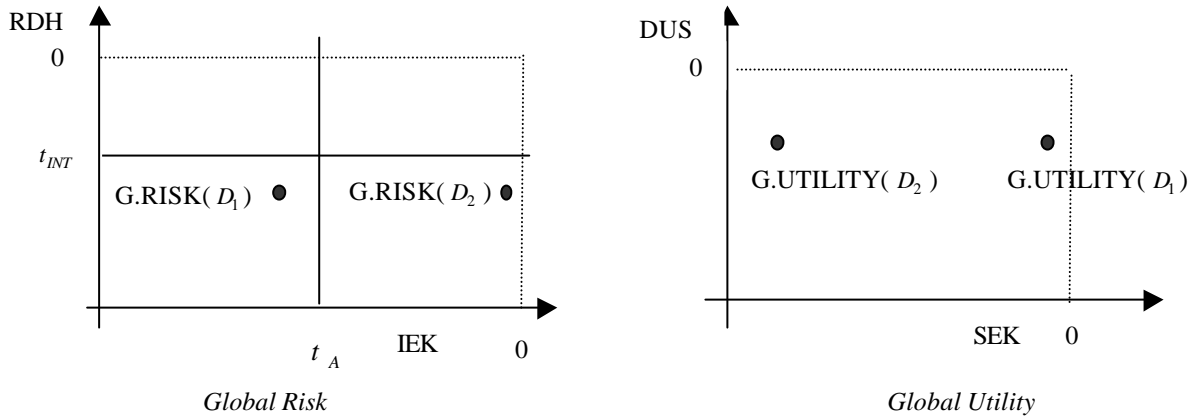


Figure 2: Trade-off of gain versus risk

29. The data sets  $D_1$  and  $D_2$  have the same risk of discredit harm (-1). Based on the definition this means that either the agency releases  $D_1$  or the agency releases  $D_2$ , the intruder is equally likely to discredit the agency claiming that confidential information has been disclosed. However the intruder's estimated knowledge when  $D_1$  is released is much smaller than the intruder's estimated knowledge when  $D_2$  is released, i.e. from the agency's point of view the release of  $D_2$  leads to a much better intruder's inference than the release of  $D_1$ . Similarly  $D_1$  and  $D_2$  have the same data utility from society's point of view (-2), i.e. either  $D_1$  is released or  $D_2$  is released the society's uncertainty about the target  $\Theta_{SOC}$  is the same. However the society's estimated knowledge when  $D_1$  is released is much higher than the society's estimated knowledge when  $D_2$  is released. This means that, from the agency's point of view, society's inference when  $D_1$  is released is much better than society's inference when  $D_2$  is released. Using threshold values  $t_{INT} = -0.8$  and  $t_A = -40$  there is only a safe data set,  $D_1$ , and therefore according with the optimality criterion  $D_1$  should be released. Actually, as we expect,  $D_1$  is always the optimal form of data release, no matter which threshold values we use but for the case in which no safe data set exists and therefore no data set should be released.

## V. CONCLUSIONS

30. This paper presents a decision-theoretic framework to measure the extent of disclosure and the data utility. The approach is innovative because it takes into account not only the intruder's and society's uncertainties about their targets but also the agency's knowledge of these targets. The new measure of disclosure discussed in the paper is the result of a different interpretation of the definition of *inferential disclosure* currently used in statistical confidentiality: "If the release of the statistic  $S$  makes it possible to determine the (microdata) values more accurately than it possible without the access to  $S$ , a disclosure has taken place...".

31. The definition was first introduced by Dalenius (1977) and then recommended by the Subcommittee on Disclosure Avoidance Techniques (1978). Although widely accepted, the definition contains many elements of ambiguity. In particular it is not clear how "more accurately" should be interpreted. In the recent

research of confidentiality the intruder's point of view is implicitly assumed. Based on the idea that the bigger is the uncertainty the smaller is the accuracy, current measures of disclosure are defined as decreasing functions of the intruder's uncertainty. As in Duncan and Lambert (1986) disclosure takes place if the uncertainty in the intruder's posterior distribution is smaller than the uncertainty in the intruder's prior distribution, i.e. if the knowledge gain is positive. Using the terminology introduced here, the recent research on confidentiality only takes into account the risk of discredit harm. The approach proposed here is also based on the definition of inferential disclosure, however distinction is made between the intruder's and the agency's perspectives. The resulting measure of disclosure is a vector whose components are the risk of discredit harm and the intruder's estimated knowledge. The risk of discredit harm is a measure of the accuracy of the intruder's inference from the intruder's point of view. The intruder's expected knowledge is instead a measure of the accuracy of the intruder's inference from the agency's point of view. The measure of the data utility is a natural extension of the measure proposed for the extent of disclosure. Again the global utility is represented as a vector whose component reflects the accuracy of the society's inference from the society's and the agency's point of view.

32. As shown in example 1 when a choice of the best form of data release is required the agency's decision based on the traditional measures of disclosure could be quite different from the agency's decision based on the global risk and global utility proposed here. I expect to apply the measure of global risk and global utility in real problems of different sizes. This future analysis will be very helpful to understand if and to what extent the global risk agrees with the measures of disclosure currently used in statistical confidentiality and, therefore, the impact of the measure of disclosure and data utility proposed in this paper in real problems.

## Acknowledgement

Preparation of this paper was supported in part by a Marie Curie Fellowship of the European Community programme "Improving The Human Research Potential" under the contract number HPMFCT-2000-00463; in part by the National Science Foundation under Grant EIA-9876619 to the National Institute of Statistical Sciences and a subcontract to Carnegie Mellon University. The content of the paper reflects the author's personal opinion. The European Commission and the National Institute of Statistical Sciences are not responsible for any views or results expressed.

## References

- Baeyens, Y. and Defays, D. (1998)**, Estimation of variance loss following from microaggregation by individual ranking method, in *Statistical Data Protection, Proceeding of the Conference*, Lisbon, Portugal: Eurostat, 101-108.
- Bernardo, J.M. and Smith, A.F.M. (1994)**, *Bayesian Theory*, Wiley.
- Dalenius, T. (1977)**, Towards a methodology for statistical disclosure control, *Statistik Tidschrift* **5**, 429-444.
- De Wall, A.G. and Willseborg, L.C.R.J. (1998)**, Optimal local suppression in microdata, *Journal of Official Statistics* **14**, 421-435.
- DeGroot, M.H. (1962)**, Uncertainty, information and sequential experiments, *Annals of Mathematical Statistics* **3**, 404-419.
- Disclosure Avoidance Techniques, S. (1978)**, *Statistical Working Paper 2, Report on Statistical Disclosure and Disclosure-Avoidance Techniques*, Federal Committee on Statistical Methodology, Office of Federal Policy and Standards, U.S. Department of Commerce, Washington D.C.
- Duncan, G.T. and Keller-McNulty, S. (1999)**, Risk of statistical confidentiality disclosure: A preliminary comparison of masked and synthetic data release, in *Workshop on Confidentiality*, National Academy of Sciences, Committee on National Statistics, Washington D.C.
- Duncan, G.T. and Lambert, D. (1986)**, Disclosure-limited data dissemination, *Journal of the American Statistical Association* **81**, 10-28.
- Duncan, G.T. and Pearson, R.B. (1991)**, Enhancing access to microdata while protecting confidentiality: prospects for the future (with discussion), *Statistical Science* **6**, 219-239.

- Duncan, G.T., Jabine, T.B., and De Wolf, V.A. (Eds.) (1993)**, *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, National Academy Press, Washington D.C. (Panel on Confidentiality and Data Access, Committee on National Statistics).
- Fienberg, S.E. (2000)**, Confidentiality and data protection through disclosure limitation: Evolving principles and technical advances (working paper).
- Fienberg, S., Makov, U.E., and Sanil, A.P. (1997)**, A Bayesian approach to data disclosure: Optimal intruder behaviour for continuous data, *Journal of Official Statistics* **13**, 75-90.
- Hurkens, C.A.J. and Tiourine, S.R. (1998)**, Models and methods for the microdata protection problem, *Journal of Official Statistics* **14**, 437-447.
- Kamlet, M.S., Klepper, S., and Frank, R.G. (1985)**, Mixing micro and macro data: Statistical issues and implication for data collection and reporting, in *Proceedings of the 1983 Public Health Conference on Records and Statistics*, U.S. Department of Health and Human Services, Hyattsville, Md.
- Mateo-Sanz, J.M. and Domingo-Ferrer, J. (1998)**, A method for data-oriented multivariate microaggregation, in *Statistical Data Protection, Proceedings of the Conference*, Lisbon, Portugal: Eurostat, 88-89.
- Pannekoek, J. and De Wall, T. (1998)**, Synthetic and combined estimators in statistical disclosure control, *Journal of Official Statistics* **14**, 399-410.
- Zaslavsky, A.M. and Horton, N.J. (1998)**, Balancing disclosure risk against the loss of non-publication, *Journal of Official Statistics* **14**, 411-419.