A User-Agency Model for Disclosure Limitation Problems

Mario TROTTINI [1]

Key words and phrases: Statistical Confidentiality; Disclosure Risk; Information Loss; Bayesian Decision Theory.

Abstract: A user-agency model is proposed to measure disclosure risk and information loss. The model simulates the interaction between users and agency when masked data are released. The framework is closely related to the one described by Etzioni and Kadane (1993) for the evaluation of experimental design when the design and the experiments are performed by different parties.The strength of the model is that we are able to characterize most of the current measures of disclosure risk and data utility in terms of quantities of straightforward interpretation in Bayesian decision theory, like users and agency loss functions, prior distributions etc. This gives a better understanding of the assumptions underlying current measures of disclosure risk and data utility and, at the same time, suggest how new and better measures can be derived.

1. INTRODUCTION

Data Disclosure Limitation denotes a set of techniques aimed to protect confidentiality in the release of statistical data. The problem is not trivial since protection of confidentiality should be achieved in a way which is compatible with the agency's mission of providing data users with good quality data. Data disclosure limitation usually involves a transformation, or *masking*, of the original data that produces a new data set that is finally released. Common masking techniques include, for example, adding noise, delete or aggregate observations, swapping of records, (see Duncan and Pearson 1991 for a review). In a typical disclosure scenario a class of alternative masking is available and the question of interest is : "which masking should the agency select ?".

Existing solutions to this problem are based on a measure of the harm that the release of a masked data can create to the data providers and/or to the statistical agency (*disclosure risk*), and on the assessment of the impact that the masking has on the quality of statistical analyses performed by the users (*information loss*). While there is general agreement on how to proceed once a measure of disclosure risk and a measure of information loss have been selected (see for example Duncan and Keller McNulty 2001, or Willenborg and de Waal 2001), how to define such measures is still an open problem.

In this paper we show that a fairly general class of measures of disclosure risk and information loss can be derived by a simple users-agency model. The underlying idea is that in order to assess disclosure risk and information loss the agency needs to predict the users' actions when a certain masked data is release, and then evaluate the consequences, for users and data provider, of such actions. 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 data provider. We refer to the first group as *society* and to the second group as the *intruder*. More formally, it is assumed that intruder and society are interested in unknown

[1] *ADDRESS FOR CORRESPONDENCE: Mario Trottini, Departamento de Estadística e Investigación Operativa, C/ Dr. Moliner, 50 46100, Burjassot, València, España, Tel. +34 -963864362, Fax: +34-963864735, e-mail: mario.trottini@uv.es*

quantities Θ_{INT} and Θ_{SOC} and they try to infer their target values based on the information contained in the released data. We propose to measure disclosure risk and information loss as agency's expected loss of the intruder's and society's optimal actions when the masked data are release. The expectation is taken with respect to the agency's beliefs about the users' actions and the agency's knowledge of the true value of the users' targets. The framework is closely related with the one illustrated in Etzioni and Kadane (1993) for the evaluation of experimental design when the design and the experiment are performed by different parties. The experimenter (the agency) has to choose an experiment (a masked data) taking into account that the results of the experiment will be analyzed by an evaluator (the users) other than the experimenter (the agency). There are however several differences that we discuss in Section 4.

The strength of our approach is that we are able to characterize a very large class of existing measures of disclosure risk and information loss in terms of quantity of straightforward interpretation in Bayesian decision theory, such as agency's and users' loss function, prior and posterior distributions, etc.. This characterization gives a better understanding of the assumptions underlying these measures and at the same time suggests how new and better measures can be defined. Previous works by Duncan and Lambert (1986), Lambert (1993), and Trottini (2001a), propose alternative characterizations based on simple users' models. These approaches are similar, in nature, to the one that we propose here but they apply to a much more restricted framework. In particular, as we show in Section 5, modeling only users' behavior force to define measures of disclosure within a given definition of disclosure risk. Same for information loss. In our framework, instead, we show that different choices of the agency's loss function in the agency's model can be used to characterize alternative definitions. This results in a much larger class of measures of disclosure risk and information loss that contain as special cases those discussed in Duncan and Lambert (1986), Lambert (1993), Trottini (2001a), and many others. The idea of using a users-agency model to define disclosure risk and information loss was already proposed by Trottini (2001b). However both the intruder's model and the agency's model there are very different from those that we use here. In Trottini (2001b) agency has no uncertainty about the users' actions when a masked data is released, and the agency's assessment of the consequences of the users' actions is made on an heuristic basis. Here the agency's uncertainty about the users' actions is incorporated in the model and the assessment of the consequences of these actions has theoretical foundations. Once again this results in a class of measures of disclosure risk and information loss much larger than the one discussed in Trottini (2001b).

The organization of the paper is fairly symmetric. Basic notations and definitions are introduced in Section 2. Sections 3 and 7 model the intruder's and society's behavior when data are released. In Section 4 a general class of measures of disclosure risk is derived from a model that describes how agency assesses the consequences of the intruder's action. The corresponding class of measures of information loss is presented in Section 8. In Sections 5 and 9 we illustrate with several examples the usefulness of our approach in understanding and comparing existing measures of disclosure risk and information loss and its potential for the definition of new ones. Section 10 contains final remarks and outline plans for future work.

2. NOTATION AND DEFINITIONS

We denote by D_0 the original data collected by the statistical agency and by $\tilde{D}$ the generic masked data. We assume that a class $\mathcal{D}$ of alternative masking is given and the agency has to choose the "best" masking in $\mathcal{D}$. We denote by D_{None} the masking that consists on deleting all the observations in the original data set D_0. We assume that releasing the original data and releasing no data are always possible options for the agency. This implies that $\mathcal{D}$ always contains at least D_0 and D_{None}.

In our framework disclosure risk and information loss measure the consequences for agency, users and data providers of the actions that users take when data are released. In particular we adopt the following definitions:

Definition 1 The *disclosure risk* is a measure of the extent to which the masked data makes it possible for an intruder to harm (the agency and/or the data providers);

Definition 2 The *information loss* is a measure of the loss that agency and society incur when the masked data are release.

We assume that, once the agency has evaluated the disclosure risk, then the agency will apply the following *disclosure rule*:

Agency's disclosure rule A masked data $\tilde{D}$ is considered for release if and only if the disclosure risk associated with $\tilde{D}$ is smaller than a fixed threshold t_A.

This rule is the standard practice for statistical agencies concerned with protection of confidentiality. Measures of disclosure risk that lead to the same disclosure rule are equivalent both from a practical and theoretic point of view (this is, for example, the case when two measures of disclosure are such that one can be expressed as an increasing function of the other). Several papers in the literature of statistical confidentiality deal directly with disclosure rules instead of presenting specific measures of disclosure risk. In the rest of the paper we discuss measures of disclosure or disclosure rules as it is convenient.

3. MODELING THE INTRUDER'S BEHAVIOR

Let $\pi^{(I)}(\cdot)$ be the intruder's prior distribution for Θ_{INT} and let $\pi^{(I)}(\cdot|\tilde{D})$ be the intruder's posterior distribution after the data $\tilde{D}$ have been released. We denote by $L^{(I)}(\cdot,\cdot)$ the intruder's loss function for the estimation problem "Estimate Θ_{INT}". $L^{(I)}(e, \theta_{INT})$ represents the loss that the intruder pays for an estimate e of Θ_{INT} when $\Theta_{INT} = \theta_{INT}$. It is assumed that $L^{(I)}(\cdot,\cdot)$ is non-negative and bounded. The intruder's optimal estimate of Θ_{INT} when $\tilde{D}$ is released is

$$\hat{\theta}^{(I)}(\tilde{D}) = argmin_e \int L^{(I)}(e;\theta_{INT})\pi^{(I)}(\theta_{INT}|\tilde{D})d\theta_{INT}, \tag{1}$$

and the expected loss associated with the intruder's optimal estimate is given by:

$$U^{(I)}(\tilde{D}) = \int L^{(I)}(\hat{\theta}^{(I)}(\tilde{D});\theta_{INT})\pi^{(I)}(\theta_{INT}|\tilde{D})d\theta_{INT} \tag{2}$$

(when clear from the context the argument $\tilde{D}$ in (1) and (2) will be omitted).

Following Duncan and Lambert (1986) we refer to $U^{(I)}(\tilde{D})$ in (2) as *intruder's uncertainty*, since it can be seen as a measure of the intruder's uncertainty about the true value of his target Θ_{INT} after the data $\tilde{D}$ have been released (see DeGroot 1962). We assume that the intruder behaves accordingly with the following decision rule :

Intruder's decision rule : Let $t^{(I)}$ be a threshold value for the intruder's maximum tolerable expected loss, $t^{(I)} \in [inf\, L^{(I)}(\cdot,\cdot),\, sup\, L^{(I)}(\cdot,\cdot)]$,

- If $U^{(I)} \le t^{(I)}$ then the intruder uses the optimal estimate $\hat{\theta}^{(I)}$, i.e. the intruder's estimates Θ_{INT} by $\hat{\theta}^{(I)}$. We denote such an action by $a_1^{(I)}$;
- If $U^{(I)} > t^{(I)}$ then the intruder does not use the optimal estimate $\hat{\theta}^{(I)}$, i.e. the intruder does not try to infer the target value Θ_{INT}. We denote such an action by $a_0^{(I)}$.

The idea underlying the intruder's decision rule is simple. The intruder takes action if and only if his uncertainty about the true value of the target Θ_{INT} is below a fixed threshold.

4. MODELING THE AGENCY'S BEHAVIOR

Let $L^{(A)}_{\Theta_{INT}}(\cdot,\cdot)$ be the agency's loss function and let IA denote the intruder's action. In our framework $L^{(A)}_{\Theta_{INT}}(a,\theta_{INT})$ represents the loss that agency and data providers pay when the intruder takes action a and the true value of the target is θ_{INT}. We assume that the loss associated with "intruder takes no action" is a constant c_0 independent of the true value of Θ_{INT} and that for the agency and the data provider "intruder takes no action" is preferred to "intruder tries to estimate Θ_{INT}" regardless of how good or how bad is the intruder's estimate $\hat{\theta}^{(I)}$ associated with $a_1^{(I)}$, i.e.

Assumption 1: $c_0 = L^{(A)}_{\Theta_{INT}}(a_0^{(I)},\theta_{INT}) \leq L^{(A)}_{\Theta_{INT}}(a_1^{(I)},\theta_{INT}) \qquad \forall\theta_{INT}, \forall a_1^{(I)}, \quad c_0 \in R.$

Assumption 1 formalizes a common paradigm in the practice of data disclosure limitation. The ideal scenario for the agency and the data providers, in terms of disclosure risk, is the one in which the intruder does not try to make any inference.

If the agency knew that the intruder's action is a (i.e. $IA = a$), and the true value of the intruder's target is θ_{INT} (i.e. $\Theta_{INT} = \theta_{INT}$), then we would have (see definition 1):

$$Disclosure\ Risk = L^{(A)}_{\Theta_{INT}}(a,\theta_{INT}). \tag{3}$$

However in most of the cases both the true values of IA and Θ_{INT} are unknown to the agency. More formally, we assume that IA and Θ_{INT}, from the agency's perspective, are random variables, and we propose to measure disclosure risk averaging (3) with respect to the agency's joint posterior distribution for IA, Θ_{INT} given the original data D_0,

$$Disclosure\ Risk = E_{IA,\Theta_{INT}|D_0}\{L^{(A)}_{\Theta_{INT}}(IA,\Theta_{INT})\}. \tag{4}$$

The measure of disclosure in (4) is closely related with the evaluation of experimental design introduced by Etzioni and Kadane (1993) for the case in which the design and the estimation are performed by different parties. In Etzioni and Kadane (1993) an experimenter (the agency) has to choose the optimal experiment (masking) in a class $\mathcal{D}$ of possible experiments (masking). The choice has to be made taking into account that an evaluator (intruder) other than the experimenter (the agency) will use the results of the experiment to try to estimate an unknown quantity Θ_{INT}. The loss associated to a certain experiment (form of data release) $\tilde{D}$ is defined as the experimenter's (agency's) expected loss of the evaluator's (intruder's) optimal estimate of Θ_{INT} when $\tilde{D}$ is performed (released). The difference between the two problems is that in our framework the posterior distributions of the experimenter (agency) and the evaluator (the intruder) are different not only because they use different priors but also because they update their prior information based on different data. The agency observes D_0 while the intruder sees only the released data $\tilde{D}$. In addition the evaluator's target is only partially known to the experimenter and the evaluator can decide whether or not he will use his optimal estimate. Finally in Etzioni and Kadane the result of the experiment is unknown a priori while in our framework $\tilde{D}$ is a trivial experiment that produces $\tilde{D}$ with probability one. As a result we do not average expression (3) over the agency's marginal distribution of $\tilde{D}$ since $\tilde{D}$ is given. Despite these differences, however, the connection between the two problems is evident. In both cases the key feature of the experimenter's evaluation model is that the results of the experiment are evaluated by decision makers other than the experimenter that produces the data.

We now state a set of assumptions that formalize what we suppose the agency knows about the intruder's action and the intruder's target and then, based on these assumptions, we derive the joint posterior distribution of IA, Θ_{INT} and we give an explicit formula to compute the disclosure risk in (4).

Assumption 2: We assume that the agency does not know the intruder's target but it knows only that the intruder's target belongs to a class $\mathcal{T}^{(I)}$ of $k \geq 1$ possible targets, i.e.

$$\Theta_{INT} \in \mathcal{T}^{(I)} = \{\Theta^1_{INT}, \ldots, \Theta^k_{INT}\}.$$

The agency in most of the cases can identify what are the possible targets for an intruder. Assumption 2 is realistic in many real problems, although the value of k can be very large.

Let J be an indicator variable such that $J = j$ if and only if $\Theta_{INT} = \Theta^j_{INT}$ and let $P(J = j)$ and $P(J = j|D_0)$ be the agency's prior and posterior probabilities for the event $J = j$. We denote by $\pi_j^{(A)}(\cdot)$ and $\pi_j^{(A)}(\cdot|D_0)$ the agency's prior and posterior distributions for Θ_{INT} conditionally on $J = j$ (i.e. $\pi_j^{(A)}(\cdot)$ and $\pi_j^{(A)}(\cdot|D_0)$ are the agency's prior and posterior distributions for Θ^j_{INT}). It is important to notice the double nature of the agency's uncertainty about the intruder's target. The agency ignores what is the target (i.e. J), and given the target what is the true value of the target (i.e. Θ^j_{INT}). We denote by $L_j^{(I)}(\cdot,\cdot)$, $\pi_j^{(I)}(\cdot)$, and $\pi_j^{(I)}(\cdot|\tilde{D})$, the intruder's loss function and the intruder's prior and posterior distribution for Θ_{INT} when $J = j$ (these were the intruder's loss function, prior and posterior distributions if the intruder's target was Θ^j_{INT}).

Assumption 3: We assume that both the intruder's priors $\pi_j^{(I)}(\cdot)$ and the intruder's loss functions, $L_j^{(I)}(\cdot,\cdot)$ are known to the agency and therefore given $J = j$, the agency knows the intruder's posterior distribution, the intruder's optimal estimate of Θ_{INT}, $\hat{\theta}_j^{(I)}$, and the intruder's uncertainty, $U_j^{(I)}$.

Assumption 3 is not reflective of real-world settings. Here we made it largely for convenience although the results that we present can be extended to the most realistic scenario where, for any given target Θ^j_{INT}, classes of intruder's prior distributions for Θ^j_{INT}, and classes of loss functions are considered instead of a single intruder's prior or a single intruder's loss function.

We denote by $t_j^{(I)}$ the intruder's threshold for the maximum tolerable expected loss when $J = j$, $t_j^{(I)} \in [inf\, L_j^{(I)}(\cdot,\cdot),\ sup\, L_j^{(I)}(\cdot,\cdot)]$. From the *intruder's decision rule* it follows that given $J = j$,

- If $U_j^{(I)} \leq t_j^{(I)}$ then the intruder uses the optimal estimate $\hat{\theta}_j^{(I)}$, i.e. the intruder's estimates Θ_{INT} by $\hat{\theta}_j^{(I)}$. We denote such an action by $a_{1j}^{(I)}$;
- If $U_j^{(I)} > t_j^{(I)}$ then the intruder does not use the optimal estimate $\hat{\theta}_j^{(I)}$, i.e. the intruder does not try to infer the target value Θ_{INT}. We denote such an action as $a_{0j}^{(I)}$.

We assume that $t_j^{(I)}$ is fixed by the intruder but unknown to the agency. More formally we assume the following:

Assumption 4a: Conditionally on $J = j$, from the agency's perspective, the intruder's threshold is a random variable, that we denote by $T_j^{(I)}$. We denote by $\pi_{T_j^{(I)}}(\cdot)$ and $\pi_{T_j^{(I)}}(\cdot|D_0)$ the agency's prior and posterior distribution for $T_j^{(I)}$;

Assumption 4b: $T_j^{(I)}$ and Θ^j_{INT} are independent;

Assumption 4c: Agency's prior and posterior distributions for $T_j^{(I)}$ are the same,

$$\pi_{T_j^{(I)}}(\cdot|D_0) = \pi_{T_j^{(I)}}(\cdot) \quad \forall j.$$

It seems reasonable to assume that the agency ignores the intruder's threshold (assumption 4a). Again classes of prior distributions can be used instead of a single distribution to formalize the agency's beliefs about the true value of $T_j^{(I)}$. Assumptions 4b and 4c are more controversial. Assumption 4b says that conditionally on $J = j$ the agency's beliefs about the intruder's threshold do not depend on the agency's beliefs about the true value of Θ_{INT}. Assumption 4c says that the original data do not contain any information about the intruder's threshold. An example might help to clarify. Suppose that the agency believes that the intruder's target is to estimate the value of a sensitive variable Θ_{INT} for a particular respondent and that the intruder is more willing to take an action when he believes that the target Θ_{INT} takes an "extreme value". Consider two scenarios: (A) the agency knows that the true value of Θ_{INT} is "extreme" and that the intruder's posterior distribution is $\pi^{(I)}(\cdot|\tilde{D})$; (B) the agency knows that the true value of Θ_{INT} is not "extreme" and the intruder's posterior distribution is $\pi^{(I)}(\cdot|\tilde{D})$ the same that in (A). What will be the agency's distribution for the intruder's threshold in (A) and (B) ? It seems reasonable to assume that in both cases the agency's beliefs about the intruder's threshold are the same. That's exactly what we would have under assumptions 4b and 4c. The underlying idea is that the intruder might fix his threshold on the basis of what he knows about Θ_{INT} but in any case he does not fix his threshold on the basis of the agency's knowledge of Θ_{INT}.

From assumptions 3 and 4 it follows that conditionally on $J = j$ the intruder's action, from the agency's perspective, is a random variable that takes values $a_{1j}^{(I)}$ and $a_{0j}^{(I)}$. We denote by $\{Pr(a_{oj}^{(I)}), Pr(a_{1j}^{(I)})\}$ and by $\{Pr(a_{oj}^{(I)}|D_0), Pr(a_{1j}^{(I)}|D_0)\}$ the agency's prior and posterior distributions of IA given $J = j$. From assumption 4 we have that: (i) these distributions depend only on the agency's distribution $\pi_{T_j^{(I)}}(\cdot)$ for the intruder's threshold, (ii) conditionally on $J = j$ prior and posterior distributions of IA are the same, i.e.:

$$\begin{aligned} Pr(a_{1j}^{(I)}|D_0) &= Pr(a_{1j}^{(I)}) = Pr(U_j^{(I)} < T_j^{(I)}) = \int_{U_j^{(I)}}^{\infty} \pi_{T_j^{(I)}}(t)dt; \\ Pr(a_{0j}^{(I)}|D_0) &= Pr(a_{0j}^{(I)}) = 1 - Pr(a_{1j}^{(I)}). \end{aligned} \tag{5}$$

We are now able to derive a formula for the disclosure risk. From (4) we have:

$$\begin{aligned} & \textit{Disclosure Risk} = E_{IA,\Theta_{INT}|D_0}\{L_{\Theta_{INT}}^{(A)}(IA, \Theta_{INT})\} \\ = \; & E_{J|D_0}[E_{IA,\Theta_{INT}|J,D_0}\{L_{\Theta_{INT}}^{(A)}(IA, \Theta_{INT})\}] \\ = \; & E_{J|D_0}[E_{\Theta_{INT}|J,D_0}[E_{IA|J,D_0}\{L_{\Theta_{INT}}^{(A)}(IA, \Theta_{INT})\}]] \\ = \; & \sum_{j=1}^{k} P(J=j|D_0) \cdot [Pr(a_{1j}^{(I)}|D_0) \cdot E_{\Theta_{INT}|J=j,D_0}\{L_{\Theta_{INT}}^{(A)}(a_{1j}^{(I)}, \Theta_{INT})\} + \\ & Pr(a_{0j}^{(I)}|D_0) \cdot E_{\Theta_{INT}|J=j,D_0}\{L_{\Theta_{INT}}^{(A)}(a_{0j}^{(I)}, \Theta_{INT})\}] \\ = \; & \sum_{j=1}^{k} P(J=j|D_0) \cdot [Pr(a_{1j}^{(I)}) \cdot E_{\Theta_{INT}|J=j,D_0}\{L_{\Theta_{INT}}^{(A)}(a_{1j}^{(I)}, \Theta_{INT})\} + Pr(a_{0j}^{(I)}) \cdot c_0], \end{aligned} \tag{6}$$

where the third last equality follows noticing that, from assumption 4b, IA and Θ_{INT} are independent given J.

Formula (6) has a straightforward interpretation. Conditionally on $J = j$ is:

$$\textit{disclosure risk} \,|\, J = j \; = Pr(a_{1j}^{(I)}) \cdot E_{\Theta_{INT}|J=j,D_0}\{L_{\Theta_{INT}}^{(A)}(a_{1j}^{(I)}, \Theta_{INT})\} + Pr(a_{0j}^{(I)}) \cdot c_0. \tag{7}$$

The conditional disclosure risk in (7) is a weighted average of the agency's and data provider's expected loss when the intruder uses his optimal estimate $\hat{\theta}^{(I)}$,

$$E_{\Theta_{INT}|J=j,D_0}\{L_{\Theta_{INT}}^{(A)}(a_{1j}^{(I)}, \Theta_{INT})\}, \tag{8}$$

and the agency's and data provider's expected loss when the intruder desists from trying to estimate Θ_{INT},

$$E_{\Theta_{INT}|J=j,D_0}\{L_{\Theta_{INT}}^{(A)}(a_{0j}^{(I)}, \Theta_{INT})\} = c_0. \tag{9}$$

The weights are the agency's posterior probabilities of the intruder's actions $a_{1j}^{(I)}$ and $a_{0j}^{(I)}$ respectively. The disclosure risk in (6) is obtained averaging the conditional disclosure risks in (7) with respect to the agency's posterior distribution for J, the indicator function that define the intruder's target. Note that in (7) the expected loss associated with action $a_{ij}^{(I)}$, $i = 0, 1$, is evaluated on the basis of the agency's knowledge of the true value of the intruder's target (both expectations in (8) and (9) are taken with respect to the agency's posterior distribution for Θ_{INT}^{j}) and on the basis of the agency's perception of the consequences of the intruder's action (in both (8) and (9) agency's loss function $L^{(A)}(\cdot, \cdot)$ is used).

In most of the examples that we consider is $c_0 = 0$ and the risk of disclosure is just the product of the probability that the intruder uses his optimal estimate times a term that measure the consequences, for the agency and the data providers, of the intruder's optimal estimate,

$$disclosure\ risk = \sum_{j=1}^{k} P(J=j|D_0) \cdot [Pr(a_{1j}^{(I)}) \cdot E_{\Theta_{INT}|J=j,D_0}\{L_{\Theta_{INT}}^{(A)}(a_{1j}^{(I)}, \Theta_{INT})\}]. \tag{10}$$

When $k = 1$ we will omit the sub-index j in both (6) and (10).

The appealing feature of the disclosure risk in (6) is that measures of disclosure (and disclosure rules) can be characterized by specific intruder's and agency's loss functions and by a specific choice for the agency's distributions for the intruder's threshold for the maximum extent of disclosure. Each of these terms has a straightforward interpretation. The agency's distributions for the intruder's threshold, $\pi_{T_j^{(I)}}(\cdot)$, formalize the agency's beliefs about the intruder's action (these distributions determine the probability that the intruder takes action $a_{ij}^{(I)}$, $i = 0, 1$). The intruder's loss function, $L^{(I)}(\cdot,\cdot)$, formalizes a precise assumption about the intruder's behavior and how the intruder will use the released data. The agency's loss function, $L_{\Theta_{INT}}^{(A)}(\cdot,\cdot)$, instead, formalizes a precise assumption about the way that agency and data providers perceive disclosure. As we show in the next Section, different definitions of disclosure of common use in statistical confidentiality correspond to different choices of the agency's loss function $L_{\Theta_{INT}}^{(A)}(\cdot,\cdot)$, and for a given definition of disclosure (i.e. for a given $L_{\Theta_{INT}}^{(A)}(\cdot,\cdot)$) well known measures of disclosure risk can be derived from (6) for specific choices of the intruder's loss function $L^{(I)}(\cdot,\cdot)$, and agency's distribution for the intruder's threshold $\pi_{T_j^{(I)}}(\cdot)$. This characterization is very useful in understanding important features of many measures of disclosure currently used in the research on statistical confidentiality and, at the same time, it suggests a natural framework for the definition of new measures.

5. EXAMPLES

In this Section we show that several measures of disclosure that have been proposed in the literature of statistical confidentiality and that are currently used by statistical agencies, belong to the general class of measures of disclosure in (6). The set of examples that we present is not intended to be exhaustive. The aim is rather to illustrate the flexibility of our approach, and its potential for the definition of new measures. We have tried to group examples by the definition of disclosure that they refer to.

5.1 Disclosure as Intruder's Uncertainty

Duncan and Lambert (1986) suggest to measure the extent of disclosure as a decreasing function of the intruder's uncertainty. The underlying idea is that the bigger is the intruder's uncertainty about the true value of Θ_{INT} after the data have been released, the smaller is the risk that an intruder can use the released data to harm the data providers or to discredit the statistical agency. The resulting disclosure rule is:

Agency's disclosure rule: A data set $\tilde{D}$ is considered for release if and only if

$$U^{(I)}(\tilde{D}) > t', \tag{11}$$

where t' is a fixed threshold.

It can be shown that (11) is a special case of the disclosure rule in (6) if we assume:

$$\begin{array}{ll} \diamond & k = 1; \\ \diamond & \Theta_{INT}\ , L^{(I)}(e, \theta_{INT})\ \pi_T^{(I)}(\cdot),\ \pi^{(A)}(\cdot) \text{ arbitrary;} \\ \diamond & c_0 = 0; \\ \diamond & L_{\Theta_{INT}}^{(A)}(a_1^{(I)}, \theta_{INT}) = c,\ \ c \geq 0\ \ \forall \theta_{INT}, \forall a_1^{(I)}. \end{array} \tag{12}$$

In particular if we take:

$$\diamond\ \pi_{T^{(I)}}(t) = Uniform[0, U^{(I)}(D_{None})], \quad c = U^{(I)}(D_{None}), \tag{13}$$

than we obtain Duncan and Lambert's *(intruder's) knowledge gain* measure of disclosure:

$$(intruder's)\ knowledge\ gain = [U^{(I)}(D_{None}) - U^{(I)}(\tilde{D})]; \tag{14}$$

while for

$$\diamond\ \pi_{T^{(I)}}(t) = Uniform[0, U^{(I)}(D_{None})], \qquad c = 1, \tag{15}$$

we obtain Duncan and Lambert's *(intruder's) relative knowledge gain* measure of disclosure:

$$(intruder's)\ relative\ knowledge\ gain = [U^{(I)}(D_{None}) - U^{(I)}(\tilde{D})]/U^{(I)}(D_{None}). \tag{16}$$

In the literature of statistical confidentiality Duncan and Lambert's approach represents, so far, the most successful attempt to provide a decision-theoretic framework to measure the extent of disclosure. The authors show that several measures of disclosure that have been proposed on an ad-hoc basis and that are currently used by statistical agency can be obtained as special cases of (11) for suitable choices of the intruder's loss function $L^{(I)}(\cdot,\cdot)$. Disclosure measures are characterized by a specific intruder's behavior when data are released. Our framework goes a step further, we model not only the intruder's behavior but also the agency's behavior. The disclosure rule in (11) as the (intruder's) knowledge gain and the (intruder's) relative knowledge gain measures of disclosure that have been proposed on an heuristic basis by Duncan and Lambert correspond in our framework to a specific choice of the agency's loss function $L^{(A)}(\cdot,\cdot)$, that formalizes how the agency perceives disclosure, and by specific choices of the agency's distribution for the intruder's threshold, $\pi_{T^{(I)}}(\cdot)$, that formalize the agency's uncertainty about the intruder's action. In particular note that in Duncan and Lambert's interpretation of disclosure, the release of a data set $\tilde{D}$ can create harm to the agency and the data providers only if the release of $\tilde{D}$ makes the intruder very confident about his inference. In this interpretation, disclosure is what the intruder believes has been disclosed from the released data, but no distinction is made between intruder's correct and incorrect inferences. In our framework this corresponds to assume that the agency's loss function $L^{(A)}(\cdot,\cdot)$ is such that $L^{(A)}_{\Theta_{INT}}(a_1^{(I)}, \theta_{INT}) = constant$, i.e. the consequences of the intruder using his optimal estimate $\theta^{(I)}$ are the same whatever is $\theta^{(I)}$ and whatever is the true value of Θ_{INT}.

Here we revisit some of the examples discussed in Duncan and Lambert (1986). Other examples of measures of disclosure based on intruder's uncertainty in Lambert (1993) are discussed as well. Disclosure rules or measures of disclosure risk are considered as it is convenient.

5.1.1 Measures of Disclosure for Tabular Data

Suppose that the released data $\tilde{D}$ consists of a contingency table with m categories. Let p_i be the tabular proportion of subjects that belongs to category i. Two common disclosure rules are:

disclosure rule 1 : Release the data if and only if

$$max_i\ p_i < t_1$$

disclosure rule 2 : Release the data if and only if

$$max_i\ p_i < t_1, \quad and \quad min_i\ p_i > t_2$$

where the thresholds t_1 and t_2 are fixed constants in (0,1).

Both rules were proposed on an ad-hoc basis by Cassel (1975) and by the Subcommittee on Disclosure Avoidance Techniques (1978) respectively. They are also discussed in Duncan and Lambert (1986). The first rule allows the release of tables only when the tabular proportions are sufficiently small. The second rule requires tabular proportions be bounded away from 0 and 1. They can be obtained as special case of (6) assuming:

- $\diamond$ Θ_{INT} = Category to which a particular respondent belongs; (17)
- $\diamond$ $$L^{(I)}(e,\theta_{INT}) = \begin{cases} 0 & \text{if } e = \theta_{INT} \\ \gamma(p_e) & \text{if } e \neq \theta_{INT} \end{cases} = L_1^{(I)}(e,\theta_{INT}) \qquad \text{for rule 1;}$$

  where $\gamma(\cdot)$ is a nonincreasing function;
- $\diamond$ $$L^{(I)}(e,\theta_{INT}) = \begin{cases} 0 & \text{if } e = \theta_{INT} \\ min(1-p_e, bp_e) & \text{if } e \neq \theta_{INT} \end{cases} = L_2^{(I)}(e,\theta_{INT}) \quad b \geq 0 \qquad \text{for rule 2;}$$
- $\diamond$ All the rest as in (12).

The first loss function in (17), $L_1^{(I)}(\cdot,\cdot)$ says that the intruder loss for incorrectly specifying that the target respondent belongs to category i is more serious when category i is rare than when it is common. The loss function $L_2^{(I)}(\cdot,\cdot)$ is inspired by a similar idea but it includes an extra penalty for incorrectly specifying a prevalent category. This penalty formalizes the loss due to the missed opportunity to identify a rare attribute of the target individual. For a detailed discussion of the two loss functions $L_i^{(I)}(\cdot,\cdot)$, $i = 1, 2$ see Duncan and Lambert (1986).

5.1.2 Measures of Disclosure for Microdata

Suppose that the original data set, D_0, is a microdata with N respondents $I_1, \ldots, I_N$ and $s+p+q$ variables, X_i, $i = 1, \ldots s+p+q$ such that:

- $X_1 \ldots, X_s$ are identifying variables, such as *Name*, *Address*, *Social Security Number*, etc.;
- $X_{s+1}, \ldots, X_{s+p}$ are key variables, i.e. can be used by an intruder to re-identify one of the respondents, linking the information about $X_{s+1}, \ldots, X_{s+p}$ contained in released data with the information contained in an external data file;
- The remaining variables, $X_{s+p+1}, \ldots, X_{s+p+q}$ are sensitive variable, i.e. they carry confidential information about the data providers.

A common form of data release in this case consists of a new microdata, $\tilde{D}$, that contains a modified version of the $p+q$ variables $X_{s+1}, \ldots, X_{s+p+q}$ (identifiers of course are not released). We denote by R_i the i^{th} record in $\tilde{D}$, and by p_i^h the intruder's posterior probability that R_i belongs to respondent I_h given $\tilde{D}$, $i, h = 1, \ldots, N$. Three common measures of disclosure risk in this case are (see Lambert, 1993):

Risk 1= $max_i\, p_i^h$, $h \in \{1, \ldots, N\}$,

Risk 2=$N^{-1} \cdot \sum_{h=1}^{N} max_i\, p_i^h$,

Risk 3 = % respondents identified

where for Risk 3 a fixed threshold t^* is considered and a respondent I_h is *identified* if and only if $max_i p_i^h > t^*$.

Risk 1 is a measure of disclosure of the type described by Duncan and Lambert (1986). It measures the risk that a particular respondent can be identified (i.e. it is a measure of *risk per record*) and can be obtained from (6) assuming:

◇ The intruder's target is to identify a particular respondent, say I_h, in $\tilde{D}$; (18)

◇ $L^{(I)}(e, \theta_{INT}) = \begin{cases} 0 & \text{if } e = \theta_{INT} \\ 1 & \text{if } e \neq \theta_{INT} \end{cases}$;

◇ $\pi_{T^{(I)}}(\cdot) = \text{Uniform}(0,1)$;

◇ $c = 1$;

◇ All the rest as in (12).

Risk 2 and Risk 3 measure the overall risk of re-identification associated with the release of $\tilde{D}$. They are obtained from the general formula for disclosure risk (6) assuming:

◇ Θ_{INT}, c as in Risk 1; (19)

◇ $k = N$;

◇ Given $J = j$ the intruder's target is to identify the record in $\tilde{D}$ that belongs to I_j;

◇ $P(J = j) = P(J = j | D_0) = 1/N\ \forall j$;

◇ $L_j^{(I)}(e, \theta_{INT}) = \begin{cases} 0 & \text{if } e = \theta_{INT} \\ 1 & \text{if } e \neq \theta_{INT} \end{cases} \quad \forall j$

◇ $\pi_{T_j^{(I)}}(\cdot) \sim \text{Uniform}(0,1)\ \ \forall j$ (for Risk 2);

◇ $\pi_{T_j^{(I)}}(\cdot) \sim \text{degenerate at } t^*\ \ \forall j$ (for Risk 3);

◇ All the rest as in (12).

Lambert (1993) refers to Risk 1 as *risk of perceived identification* (for the target I_h) and to Risk 2 and Risk 3 as *overall risk of perceived identification* since they measure the intruder's perceived probability of a correct link for a particular respondent or for the overall data set.

What distinguish the three measures of disclosure is: (i) the agency's uncertainty about the intruder's target (the distribution of J) (ii) the agency's distribution for the intruder's threshold ($\pi_{T_j^{(I)}}(\cdot)$). For Risk 1 is J degenerate at some value h while for Risk 2 and Risk 3 is J Uniform on $\{1, \ldots, N\}$. In addition $\pi_{T_j^{(I)}}(\cdot)$ is uniform for Risk 1 and Risk 2 and degenerate at the intruder's threshold t^* for Risk 3. This imply that in Risk 1 we are assuming that the agency knows exactly which individual is the intruder's target, while in Risk 2 and Risk 3 we are assuming that the agency knows exactly the intruder's threshold for the maximum tolerable expected loss. In both cases these assumptions seems quite unrealistic.

5.2 Disclosure as Intruder's Correct Inference

Many papers equate disclosure with revealing the true identity or the true attribute of a respondent (see Blien, Wirth and Muller 1992). This interpretation of disclosure is rather different from the one proposed by Duncan and Lambert (1986). In particular it is assumed that intruder's incorrect inferences about Θ_{INT} are not dangerous and only correct inferences can create harm. A common measure of disclosure of this type is the *risk of true identification* discussed by Lambert (1993),

risk of true identification = % respondents correctly identified

(using the notation in subsection 5.1, a respondent I_h is *correctly identified* if and only if $max_i p_i^h > t^*$ and $max_i p_i^h = p_{h^*}^h$, where R_{h^*} denotes the record in $\tilde{D}$ that belongs to I_h).

The defining feature of this measure of disclosure is that only correct identification is considered dangerous for the agency and the data providers while there is no harm if the intruder makes an incorrect link. Equating disclosure with intruder's correct inference in our framework is equivalent to assume that the agency's loss function $L^{(A)}_{\Theta_{INT}}(\cdot,\cdot)$ is such that $L^{(A)}_{\Theta_{INT}}(a_1^{(I)}, \theta_{INT})$ decreases as $\hat{\theta}^{(I)}$, the optimal estimate of Θ_{INT} associated with $a_1^{(I)}$, moves away from θ_{INT} (i.e. the worst is the intruder's estimate of Θ_{INT} the smaller the harm to the agency and to the data providers). A simple way to define $L^{(A)}_{\Theta_{INT}}(\cdot,\cdot)$ is the following. Let $L(\cdot,\cdot)$ be a loss function for the agency's estimation problem "Estimate Θ_{INT}" then define:

$$L^{(A)}_{\Theta_{INT}}(a, \theta_{INT}) = \begin{cases} -L(\hat{\theta}^{(I)}, \theta_{INT}) & \text{if } a = a_1^{(I)} \\ c_0 = -sup\ L(e, \theta_{INT}) & \text{if } a = a_0^{(I)} \end{cases}. \tag{20}$$

Note that (20) satisfies assumption 1 since

$$L^{(A)}_{\Theta_{INT}}(a_1^{(I)}, \theta_{INT}) = -L(\hat{\theta}^I, \theta_{INT}) \geq -sup\ L(\cdot,\cdot) = c_0 \quad \forall a_1^{(I)},\ \forall \theta_{INT},$$

as required.

Using the agency's loss function in (20) the general formula for disclosure risk in (6) becomes:

$$Disclosure\ Risk = \sum_{j=1}^{k} P(J=j|D_0) \cdot [Pr(a_{1j}^{(I)}) \cdot [-E_{\Theta_{INT}|J=j,D_0}\{L(a_{1j}^{(I)}, \theta_{INT})\}] + Pr(a_{0j}^{(I)}) \cdot c_0]. \tag{21}$$

Note that (21) behaves as we expect. If the released data are such that intruder has no uncertainty about his estimate of Θ_{INT} (i.e. $k = 1$, $U^{(I)}(\tilde{D}) = 0$) and from the agency's perspective the intruder's estimate is correct with probability 1 ($\pi^{(A)}_{\Theta_{INT}}(\cdot|D_0)$ degenerate at $\hat{\theta}^{(I)}$), then whatever is the agency's distribution for the intruder's threshold, disclosure risk achieves its maximum ($U^{(I)}(\tilde{D}) = 0$ implies $P(a_1^{(I)}) = 1\ \forall \pi_{T^{(I)}}(\cdot)$ and we have: disclosure risk $= -L(\hat{\theta}^{(I)}, \hat{\theta}^{(I)}) = -inf\ \ L(e, \theta_{INT})$). On the other hand, if the released data are such that: (i) the intruder desists from estimating Θ_{INT} ($Pr(a_{1j}^{(I)}) = 0$), or (ii) the intruder has no uncertainty about his estimate ($k = 1$, $U^{(I)}(\tilde{D}) = 0$) and from the agency's perspective intruder's estimate is very poor with probability 1 ($\pi^{(A)}_{\Theta_{INT}}(\cdot|D_0)$ degenerate at some value θ^* very "far" from $\hat{\theta}^{(I)}$)

then disclosure risk achieves its minimum (disclosure risk $= c_0$). For (i), the result is straightforward. For (ii) we have: $P(a_1^{(I)}) = 1$ and

$$Disclosure\ \ Risk = [-E_{\Theta_{INT}|J=j,D_0}\{L(a_{1j}^{(I)}, \theta_{INT})\}] = -L(\hat{\theta}^{(I)}, \theta^*) \approx -sup\ L(\cdot,\cdot) = c_0$$

(note that the range of the disclosure risk in (21) is the interval $[c_0, -inf L(\cdot,\cdot)]$ which can be re-written as $[c_0, -L(e,e)]$, $\forall e$ since $L(\cdot,\cdot)$ is a loss function for an estimation problem and takes its minimum when the estimate is equal to the true value of the parameter).

The risk of true identification is special case of (21) if we assume:

- $\diamond$ $L_{\Theta_{INT}}^{(A)}(\cdot,\cdot)$ as in (20) with $L(e,\theta_{INT}) = \begin{cases} -1 & \text{if } e = \theta_{INT} \\ 0 & \text{if } e \neq \theta_{INT} \end{cases} \quad \forall j;$ (22)
- $\diamond$ The agency knows the record in $\tilde{D}$ that belongs to each I_h, *i.e.* $\pi^{(A)}(\theta_{INT}|D_0)$ degenerate at h^* (R_{h^*} is the record in $\tilde{D}$ that belongs to I_h);
- $\diamond$ All the rest as in (19) for risk 3.

Consider for a moment the term that measures the consequences of the intruder's action $a_1^{(I)}$ in the disclosure risk in (21) (to simplify things suppose $k = 1$):

$$-E_{\Theta_{INT}|D_0}\{L(a_1^{(I)}, \theta_{INT})\}. \tag{23}$$

Denote by $\hat{\theta}^{(A)}$ and $U^{(A)}$ the agency's optimal estimate of Θ_{INT} and the agency's uncertainty with respect to the loss function $L(\cdot,\cdot)$,

$$\begin{aligned} \hat{\theta}^{(A)} &= argmin_e \int L(e;\theta_{INT})\pi^{(A)}(\theta_{INT}|D_0)d\theta_{INT}; \\ U^{(A)} &= \int L(\hat{\theta}^{(A)};\theta_{INT})\pi^{(A)}(\theta_{INT}|D_0)d\theta_{INT}. \end{aligned} \tag{24}$$

We can decompose (23) as follows:

$$-E_{\Theta_{INT}|D_0}\{L(a_{1j}^{(I)}, \theta_{INT}) = -[U^{(A)} + \Delta] \tag{25}$$

where Δ is a "distance" between intruder's and agency's optimal estimates of Θ_{INT},

$$\Delta = E_{\theta_{INT}|D_0}[L(\theta^{(I)}, \theta_{INT})] - E_{\theta_{INT}|D_0}[L(\theta^{(A)}, \theta_{INT})].$$

Δ "small" means that although sub-optimal, from the agency's point of view, $\theta^{(I)}$ leads almost to the same expected loss than $\theta^{(A)}$ (note that since $\theta^{(A)}$ by definition minimizes $E_{\Theta_{INT}|D_0}L(a,\theta_{INT})$, Δ is always non negative).

From (25) it follows that if the intruder decides to estimate Θ_{INT}, the harm to the agency and data providers is a decreasing function of the agency's uncertainty about the intruder's target and the "difference" between intruder's and agency's optimal estimate of Θ_{INT} as measured by Δ. This is coherent with the intuition. The assumption underlying the measure of disclosure in (21) is that only intruder's correct inference are harmful for the agency and the data providers. Therefore we would expect that (i) the harm deriving from the intruder using his optimal estimate $\hat{\theta}^{(I)}$ should be a decreasing function of the difference between the intruder's optimal estimate of Θ_{INT} and what the agency believes to be the true value of Θ_{INT}, and (ii) if the agency uncertainty about the true value of Θ_{INT} is high there is (almost) no harm regardless of what is the intruder's optimal estimate $\theta^{(I)}$, since whatever is $\hat{\theta}^{(I)}$, on average $\hat{\theta}^{(I)}$ is far from what the agency believes to be the true value of Θ_{INT}.

An interesting interpretation holds when agency and intruder share a quadratic loss function. The term Δ that measures the "difference" between intruder's and agency's optimal inferences in this case is nothing but the square of the Euclidean distance between intruder's and agency's optimal estimates of Θ_{INT}:

$$\Delta = (\theta^{(I)} - \theta^{(A)})^2, \tag{26}$$

where $\theta^{(I)}$ and $\theta^{(A)}$ are the expected values of Θ_{INT} under the intruder's and the agency's posterior distribution respectively. If we further assume:

$$\diamond\ \pi_{T^{(I)}}(\cdot) \sim \text{ degenerate at } sup\, L^{(I)}(\cdot,\cdot),$$

then we have $P(a_1^{(I)}) = 1\ \forall a_1^{(I)}$ and disclosure risk is a decreasing function of the agency's uncertainty and the square of the Euclidean distance between intruder's and agency's optimal estimates of Θ_{INT}:

$$Disclosure\ Risk = -[U^{(A)} + (\theta^{(I)} - \theta^{(A)})^2]. \tag{27}$$

The disclosure risk in (27) is closely related with the *risk of disclosure harm* (RDL) proposed by Trottini (2001b),

$$RDL = -[U^{(A)} + D(\pi^{(I)}_{\Theta_{INT}}(\cdot|\tilde{D}), \pi^{(A)}_{\Theta_{INT}}(\cdot|D_0))], \tag{28}$$

where $D(\cdot,\cdot)$ represents a measure of divergence between distributions.
The risk of disclosure harm was proposed by Trottini (2001b) on an ad-hoc basis. The goal was to define a component of a global measure of disclosure that could take into account the quality of the intruder's inference and that penalizes good inferences more than bad inferences. It follows from (27) that the risk of disclosure harm belongs to the general class of disclosure measures in (6) if we define the divergence $D(\cdot,\cdot)$ between two probability distributions as the square of the Euclidean distance between the expected value of the two distributions.

5.3 Disclosure as Embarrassment to the Data Providers

Suppose that a statistical agency has to decide about the release of a microdata containing a sensitive variable $Xp + q + 1$, that measures whether or not a respondent is HIV positive (assume $Xp + q + 1$ takes value 1 if a respondent is HIV positive, 0 otherwise). Suppose also that the intruder is an insurance company that wants to disclose the value of $Xp + q + 1$ for a particular respondent applying for a new policy. In this case it is reasonable to assume that no respondent wants an intruder to decide that the sensitive variable takes value 1 and a reasonable measure of disclosure is:

Risk 4= % or respondents identified as HIV positive

where as in Section 5.2 a respondent I_h is *identified* if and only if $max_i\, p_i^h > t^*$.

The above example is representative of a much larger class of examples discussed by Lambert (1993) where the released data contains a sensitive variable/vector X and no respondent wants an intruder to decide that X takes a particular value x^* regardless of what is the true value of X. This, in our framework, corresponds to assume that the intruder's target is the sensitive variable X and the agency's loss function $L^{(A)}_{\Theta_{INT}}(\cdot,\cdot)$ is such that $L^{(A)}_{\Theta_{INT}}(a_1^{(I)}, \theta_{INT})$ decreases as $\hat{\theta}^{(I)}$, the optimal estimate of Θ_{INT} associated with $a_1^{(I)}$, moves away from x^* (i.e. the "farther" is the intruder's estimate from x^* the smaller the harm to the agency and to the data providers. For example in the HIV example above, Risk 4 can be obtained from (6) assuming:

- ◇ $\Theta_{INT} =$ the value of X_{p+q+1} for respondent I_h, $h \in \{1,\ldots,N\}$; (29)
- ◇ $\Theta^j_{INT} =$ the value of X_{p+q+1} for respondent I_j, $\forall j$;
- ◇ $L_j^{(A)}(a, \theta_{INT}) = \begin{cases} 1 & \text{if } a = \theta_{INT} \\ 0 & \text{if } a \neq \theta_{INT} \end{cases} \quad \forall j;$
- ◇ The agency knows the true value of X_{p+q+1} of each respondent;
- ◇ All the rest as in (22).

Note that risk 4 is substantially different from the other measures of disclosures discussed in Sections 5.1 and 5.2. To see this suppose, for example, that the released microdata is such that the intruder's always makes a link, that the link is always correct, and that for all subjects in the released data is $Xp+q+1 = 0$ (i.e. in the released data no subject appears as HIV positive). In this case Risk 1, Risk 2 Risk 3 and the risk of true identification will take their maximum value (i.e. 1) while Risk 4 will take its minimum value (i.e. 0).

6. INFORMATION LOSS

In our framework the problem of measuring information loss is formally equivalent to the problem of measuring disclosure risk. The agency releases the data, society makes its inference about the target Θ_{SOC} based on the released data, the agency evaluates the consequences of society's inference taking into account its knowledge of society's inference and its knowledge of the true value of society's target. Here we briefly describe society's model and the agency's model (for data utility) and then we derive a general measure of information loss that mirrors the measure of disclosure risk in Section 4. Given the symmetry between the two problems we only comment models' assumptions when these somehow differ from the corresponding assumptions in Section 4. As for the risk of disclosure we illustrate with several examples the flexibility of our approach its usefulness to understand and compare existing measures of information loss, and its potential for the definition of new ones.

7. MODELING SOCIETY'S BEHAVIOR

Let $\pi^{(S)}(\cdot)$ be society's prior distribution for Θ_{SOC} and let $\pi^{(S)}(\cdot|\tilde{D})$ be society's posterior distribution after the data $\tilde{D}$ have been released. We denote by $L^{(S)}(\cdot,\cdot)$ society's loss function for the estimation problem "Estimate Θ_{SOC}". $L^{(S)}(e,\theta_{SOC})$, represents the loss that society pays for an estimate e of Θ_{SOC} when $\Theta_{SOC}=\theta_{SOC}$. We assume that $L^{(S)}(\cdot,\cdot)$ is non-negative and bounded. Society's optimal estimate of Θ_{SOC} when $\tilde{D}$ is released is

$$\hat{\theta}^{(S)}(\tilde{D}) = argmin_e \int L^{(S)}(e;\theta_{SOC})\pi^{(S)}(\theta_{SOC}|\tilde{D})d\theta_{SOC},$$

and the expected loss associated with society's optimal estimate is given by:

$$U^{(S)}(\tilde{D}) = \int L^{(S)}(\hat{\theta}^{(S)}(\tilde{D});\theta_{SOC})\pi^{(S)}(\theta_{SOC}|\tilde{D})d\theta_{SOC}.$$

Following the terminology used in Section 3, we refer to $U^{(S)}(\tilde{D})$ as *society's uncertainty*. Again when clear from the context the argument $\tilde{D}$ in the expression of $\hat{\theta}^{(S)}$ and $U^{(S)}$ will be omitted.

We assume that society behaves accordingly with the following decision rule :

Society's decision rule : Let $t^{(S)}$ be a threshold value for society's maximum tolerable expected loss, $t^{(S)} \in [inf\, L^{(S)}(\cdot,\cdot),\ sup\, L^{(S)}(\cdot,\cdot)]$

- If $U^{(S)} \leq t^{(S)}$ then society uses the optimal estimate $\hat{\theta}^{(S)}$, i.e. society estimates Θ_{SOC} by $\hat{\theta}^{(S)}$. We denote such an action by $a_1^{(S)}$;
- If $U^{(S)} > t^{(S)}$ then society does not use the optimal estimate $\hat{\theta}^{(S)}$, i.e. society does not try to infer the target value Θ_{SOC}. We denote such an action by $a_0^{(S)}$.

8. MODELING THE AGENCY'S BEHAVIOR

Let $L^{(A)}_{\Theta_{SOC}}(\cdot,\cdot)$ be the agency's loss function and let SA denote society's action. In our framework $L^{(A)}_{\Theta_{SOC}}(a,\theta_{SOC})$ represents the loss that the agency incurs when society takes action a and the true value of the target is θ_{SOC}. As in Section 4 we assume that the loss associated with "society's takes no action" is a constant, c_1, independent of the true value of Θ_{SOC}.

Assumption 5: $L^{(A)}_{\Theta_{SOC}}(a_0^{(S)},\theta_{SOC}) = c_1 \quad \forall\, \theta_{SOC}, \quad c_1 \in R.$

However in this case we do not make any assumption about the "ordering" of $a_0^{(S)}$ and $a_1^{(S)}$. While it is clear that a very good estimate of Θ_{SOC} is always preferred to "society takes no action" a poor estimate of Θ_{SOC} is not necessary preferred to "society takes no action". As we show in Section 9, it depends on the context. As a result we do not make here the equivalent of assumption 1 in Section 4.

If the agency knew that society takes action a, (i.e. $SA = a$) and that $\Theta_{SOC} = \theta_{SOC}$, then the information loss would be (see definition 2):

$$Information\ Loss = L^{(A)}_{\Theta_{SOC}}(a, \theta_{SOC}). \tag{30}$$

However as for the disclosure risk we assume that from the agency's perspective both society's action, SA, and Θ_{SOC} are random variables and we propose to measure information loss averaging (30) w.r.t. the agency's joint posterior distribution for SA, and Θ_{INT} given D_0, i.e.,

$$Information\ Loss = E_{SA,\Theta_{SOC}|D_0}[L^{(A)}_{\Theta_{SOC}}(IA, \Theta_{SOC})]. \tag{31}$$

We now state a set of assumptions that mirror assumptions 2-4 in Section 4. They formalize what we suppose the agency knows about society's action and society's target. Based on these assumptions we derive the joint posterior distribution of SA, Θ_{INT} and we give an explicit formula to compute the information loss in (31).

Assumption 6: The agency does not know society's target but it knows only that society's target belongs to a class $\mathcal{T}^{(S)}$ of $m \geq 1$ possible targets, i.e.

$$\Theta_{SOC} \in \mathcal{T}^{(S)} = \{\Theta^1_{SOC}, \ldots, \Theta^m_{SOC}\}.$$

We denote by H an indicator variable such that $H = h$ if and only if $\Theta_{SOC} = \Theta^h_{SOC}$, $h = 1, \ldots, m$. Let $P(H = h)$ and $P(H = h|D_0)$ be the agency's prior and posterior probabilities for the event $H = h$. We denote by $\pi^{(A)}_h(\cdot)$ and $\pi^{(A)}_h(\cdot|D_0)$ the agency's prior and posterior distributions for Θ_{SOC} conditionally on $H = h$. We denote by $L^{(S)}_h(a, \theta_{SOC})$, $\pi^{(S)}_h(\cdot)$, and $\pi^{(S)}_h(\cdot|\tilde{D})$, society's loss function and society's prior and posterior distribution for Θ_{SOC} when $H = h$.

Assumption 7: We assume that both society's priors $\pi^{(I)}_h(\cdot)$ and society's loss functions, $L^{(S)}_h(e, \theta_{SOC})$ are known to the agency and therefore given $H = h$, the agency knows society's posterior distribution, society's optimal estimate of Θ_{SOC}, $\hat{\theta}^{(S)}_h$, and society's uncertainty, $U^{(S)}_h$.

We denote by $t^{(S)}_h$ society's threshold for the maximum tolerable expected loss when $H = h$. From the *society's decision rule* it follows that conditionally on $H = h$,

- If $U^{(S)}_h \leq t^{(S)}_h$ then society uses the optimal estimate $\hat{\theta}^{(S)}_h$. We denote such an action by $a^{(S)}_{1h}$;
- If $U^{(S)}_h > t^{(S)}_h$ then society does not use the optimal estimate $\hat{\theta}^{(S)}_h$, i.e. society does not try to infer the target value Θ_{SOC}. We denote such an action by $a^{(S)}_{0h}$.

We assume that $t^{(S)}_h$ is fixed by the user but unknown to the agency. More formally we assume the following:

Assumption 8a: Conditionally on $H = h$, from the agency's perspective, society's threshold is a random variable, that we denote by $T^{(S)}_h$. We denote by $\pi_{T^{(S)}_h}(\cdot)$ and by $\pi_{T^{(S)}_h}(\cdot|D_0)$ the agency's prior and posterior distribution for $T^{(S)}_h$;

Assumption 8b: $T^{(S)}_h$ and Θ^h_{SOC} are independent;

Assumption 8c: Agency's prior and posterior distributions for $T^{(S)}_h$ are the same,

$$\pi_{T^{(S)}_h}(\cdot|D_0) = \pi_{T^{(S)}_h}(\cdot) \quad \forall h.$$

From assumptions 7 and 8 it follows that, conditionally on $H = h$, society's action, from the agency's perspective, is a random variable that takes values $a^{(S)}_{1h}$ and $a^{(S)}_{0h}$. We denote by $Pr(a^{(S)}_{ih})$, and $Pr(a^{(S)}_{ih}|D_0)\}$ the agency's prior and posterior probability for the event $SA = a^{(S)}_{ih}$, $i = 0, 1$. From assumption 8 we have:

$$Pr(a^{(S)}_{1h}|D_0) = Pr(a^{(S)}_{1h}) \quad = \quad Pr(U^{(S)}_h < T^{(S)}_h) = \int_{U^{(S)}_h}^{\infty} \pi_{T^{(S)}_h}(t)dt.$$

We are now able to derive a formula for the information loss. From (31) we have:

$$
\begin{aligned}
Information\ Loss &= E_{SA,\Theta_{SOC}|D_0}\{L^{(A)}_{\Theta_{SOC}}(SA,\Theta_{SOC})\} \qquad (32)\\
&= E_{H|D_0}[E_{SA,\Theta_{SOC}|H,D_0}\{L^{(A)}_{\Theta_{SOC}}(SA,\Theta_{SOC})\}]\\
&= E_{H|D_0}[E_{\Theta_{SOC}|H,D_0}[E_{SA|H,D_0}\{L^{(A)}_{\Theta_{SOC}}(SA,\Theta_{SOC})\}]]\\
&= \sum_{h=1}^{m} P(H=h|D_0)\cdot[P(a^{(S)}_{1h})\cdot E_{\Theta_{SOC}|H=h,D_0}\{L^{(A)}_{\Theta_{SOC}}(a^{(S)}_{1h},\Theta_{SOC})\}\\
&\quad +P(a^{(S)}_{0h})\cdot c_1].
\end{aligned}
$$

The interpretation of (32) is similar to the interpretation of the formula for disclosure risk in (6). Again the framework is the same than the one described in Etzioni and Kadane (1993) with agency and society playing the roles of the experimenter and the evaluator.

9. EXAMPLES

In this Section we show that many of the current measures of information loss can be derived as special case of (32). The following subsections 9.1-9.3 mirror the corresponding subsections in Section 5. As for the risk of disclosure, measures of information loss that lead to the same "ordering" of masking, are considered equivalent. This is for example the case when two measures of information loss are such that one can be expressed as an increasing function of the other.

9.1 Information Loss as Society's Uncertainty

Trottini (2001a) defines a general class of measures of information loss that mirror the measures of disclosure risk proposed by Duncan and Lambert (1986). The underlying assumption is that the bigger is society's uncertainty about the true value of Θ_{SOC} the bigger is the information loss. For example for the case in which society has already substantial knowledge about the true value of its target before the release of the data, Trottini (2001a) suggests to measure information loss as the difference or the relative difference between society's posterior and prior uncertainty,

$$Information\ Loss\ 1 = [U^{(S)}(\tilde{D}) - U^{(S)}(D_{None})], \qquad (33)$$

$$Information\ Loss\ 2 = [U^{(S)}(\tilde{D}) - U^{(S)}(D_{None})]/U^{(S)}(D_{None}). \qquad (34)$$

As in Duncan and Lambert (1986), the characteristic of these measures of information loss is that no distinction is made between correct and incorrect society's inferences. In our framework this is equivalent to assume an agency's loss function such that $L^{(A)}_{\Theta_{SOC}}(a^{(S)}_1,\theta_{SOC}) =$ constant. In particular it can be shown that (33) and (34) can be obtained from (32) if we assume:

$$
\begin{aligned}
&\diamond \quad h=1; \qquad (35)\\
&\diamond \quad \Theta_{SOC}\ , L^{(S)}_{\Theta_{SOC}}(\cdot,\cdot)\ , \pi^{(S)}_T(\cdot), \pi^{(A)}(\cdot)\ \text{ arbitrary;}\\
&\diamond \quad L^{(A)}_{\Theta_{SOC}}(a,\theta_{SOC}) = \begin{cases} -c & \text{if } a=a^{(S)}_1 \\ 0 & \text{if } a=a^{(S)}_0 \end{cases} \quad c>0;\\
&\diamond \quad \pi_{T^{(S)}}(t) = Uniform[0, U^{(S)}(D_{None})];\\
&\diamond \quad c = U^{(S)}(D_{None}) \qquad \text{for (33);}\\
&\diamond \quad c = 1 \qquad \text{for (34).}
\end{aligned}
$$

Note that although from the agency's perspective there is no difference between a good society's estimate and a poor society's estimate of Θ_{SOC} ($L^{(A)}_{\Theta_{SOC}}(a^{(S)}_1,\theta_{SOC}) =$ constant), the underlying assumption is that any society's estimate of Θ_{SOC} is preferred to "society takes no action" ($L^{(A)}_{\Theta_{SOC}}(a^{(S)}_1,\theta_{SOC}) = -c < 0 = L^{(A)}_{\Theta_{SOC}}(a^{(S)}_0,\theta_{SOC})\ \forall\theta_{SOC}$).

9.1.1 Measures of Information Loss Based on Entropy

Another class of measures of information loss that depend only on society's uncertainty is the class of measures of information loss based on entropy proposed by Willenborg and de Waal (2001). Here we illustrate a simple example where the original data consists of a frequency count table and cell suppression is applied. Similar results, however hold for other forms of data release (microdata, magnitude tables) and alternative disclosure limitation techniques (rounding, table redesign etc.).

Consider a frequency count table F with N cells (including the marginals but without the grand total that is known to be N_{TOT}). Suppose that the released data is obtained suppressing $n < N$ cells from the original table F. We denote by X_i the value of the i^{th} cell in F and by $\{X_{j_1}, \dots, X_{j_n}\}$ the set of values in F that have been suppressed, $j_i \in \{1, \dots, N\}$. Given the released table, the variables X_{j_i} define a set of linear equalities. These equality results from the additivity of the table and its marginals. Every combination of $X_{j_1}, \dots, X_{j_n}$ that satisfies the set of linear equalities is called *solution set.* We denote the solution set by S, the generic element in S by s_i and the measure of S by $M(S)$, i.e. $S = \{s_1, \dots, s_{M(S)}\}$. Willenborg and de Waal propose to measure the information loss due to suppressing cells in the original table as:

$$Information\ Loss = log M(S), \tag{36}$$

which is the entropy if we assume a uniform distribution on S (the uniform distribution in this case is quite reasonable since without additional information each solution in S is equally likely to be the right one) .

It can be shown that (36) can be obtained as special case of (32) assuming:

(i) society's target is to report the distribution of $X = (X_1, \dots, X_N)$,

$$\Theta_{SOC} = \{\Theta_1, \dots, \Theta_{M(W)}\}$$

where $M(W)$ is the measure of W, the support of X (or the space of tables spanned by $X_1, \dots, X_N$),

$$W = \{(x_1, \dots, x_N) : x_i \text{ non-negative integer and } \sum_{i=1}^{N} x_i = N_{TOT}\},$$

$\mathbf{w_i}$ denotes the generic element in W and Θ_i is the probability that $X = \mathbf{w_i}$, $i = 1, \dots, M(W)$.

(ii) $L^{(S)}_{\Theta_{SOC}}(e, \theta_{SOC}) = -\sum_{i=1}^{M(W)} log(e_i) \cdot \theta_i$ where $e = (e_1, \dots, e_{M(W)})$ is the generic estimate of Θ_{SOC}

(iii) $\pi^{(S)}_{\Theta_{SOC}}(\cdot) \sim$ Uniform on $\{\pi_1(\cdot), \dots, \pi_{M(W)}(\cdot)\}$, where $\pi_i(\cdot)$ is the distribution degenerate at $\mathbf{w_i}$ (this is equivalent to put a uniform distribution over the set of tables in W).

(iv) $\pi_{T(S)}(\cdot) \sim$ Uniform $[0, log M(W)]$ (note that $[0, log(M(W))]$ is the range of society's uncertainty $U^{(S)}(\tilde{D})$ when cell suppression is used. $U^{(S)}(\tilde{D}) = 0$ when no cell is suppressed, $U^{(S)}(\tilde{D}) = log(M(W))$ when all cells (but the grand total) are suppressed;

(v) All the rest as in (35) with $c = 1$.

Proof Under the above assumptions (32) can be rewritten as follows:

$$Information\ Loss = log(M(S))/log(M(W)) - 1$$

which is a one to one increasing function of $log(M(S))$.

9.2 Measures of Information Loss Based on Distance Between Inference with the Original Data and Inference using the Released Data

Many measures of information loss of common use in the research of statistical confidentiality are defined in terms of "distance" between inferences based on the original data and inferences based on the released data. This is the standard case when society's target is to estimate the parameters of a parametric model.

In Corsini, Franconi, Pagliuca and Seri (1998), for example, the information loss associated with alternative microaggregation masking in a econometric models is defined as:

$$Information\ Loss = |\hat{\theta}^{(S)}_{SOC}(\tilde{D}) - \hat{\theta}^{(S)}_{SOC}(D_0)|, \tag{37}$$

where $\hat{\theta}^{(S)}_{SOC}(\tilde{D})$ is society's optimal estimate of the parameters vector based on the released data, $\hat{\theta}^{(S)}_{SOC}(D_0)$ is the corresponding society's optimal estimate based on the original data and for a vector v, $|v|$ denotes the norm of v.

The assumption underlying the measures of information loss in (37) and related measures, is that the closer is the inference based on the released data to the inference based on the original data, the smaller is the information loss. In our framework this corresponds to assume that the agency's loss function $L^{(A)}_{\Theta_{SOC}}(\cdot,\cdot)$ is such that $L^{(A)}_{\Theta_{SOC}}(a^{(S)}_1, \theta_{SOC})$ increases when $\hat{\theta}^{(S)}(\tilde{D})$, the optimal estimate of Θ_{SOC} associated with $a^{(S)}_1$, moves away from $\hat{\theta}^{(S)}_{SOC}(D_0)$ (the latter in our framework is a constant that does not depend on the particular form of data release). For example, it can be shown that (37) is a special case of (32) assuming:

- ◇ $h = 1$; (38)
- ◇ $\Theta_{SOC},\ L^{(S)}_{\Theta_{SOC}}(\cdot,\cdot),\ \pi^{(A)}(\cdot)$ arbitrary;
- ◇ $\pi_T(S)$ degenerate at $sup_{\tilde{D}}(L^{(S)}_{\Theta_{SOC}}(\cdot,\cdot)$;
- ◇ $L^{(A)}_{\Theta_{SOC}}(a, \theta_{SOC}) = \begin{cases} |\hat{\theta}^{(S)}_{SOC}(\tilde{D}) - \hat{\theta}^{(S)}_{SOC}(D_0)| & \text{if } a = a^{(S)}_1 \\ arbitrary & \text{if } a = a^{(S)}_0 \end{cases}$.

Note that the extent to which society's inference is correct or precise does not matter here. Consider the following two scenarios: (A) the inference based on the released data is exactly the same than the inference based on the original data but both inferences are very poor (because society's prior is misleading for example); (B) the inference based on the released data is exactly the same than the inference based on the original data and both inferences are very good. In both (A) and (B) the measure of information loss takes its minimum although the quality of society's inference is very different in the two cases. In our framework this formalized by an agency's loss function $L^{(A)}_{\Theta_{SOC}}(\cdot,\cdot)$, that is independent on the true value of Θ_{INT}.

9.3 Information Loss as Society's Correct Inference

It seems very natural to measure information loss in terms of distances between society's optimal estimate of the target Θ_{SOC} and what the agency believes to be the true value of Θ_{SOC}. However, quite surprisingly, very few examples of such measures has been derived in the literature of statistical confidentiality. As we commented in previous sections most of the measures of information loss are based on society's uncertainty or on distance between society's inference with the original data and society's inference with the released data. These type of measures, however, do not take into account that: (i) society's inference based on the original data can be incorrect (because society's prior is misleading, for example) or inconclusive (since original data do not contain a lot of information and society's prior is non-informative), and (ii) agency's and society's posterior distributions for Θ_{SOC} given the original data D_0 can be quite different due to different priors. The interpretation of information loss as "distance" between society's estimate of Θ_{SOC} and agency's beliefs about the true value of Θ_{SOC} in our framework corresponds to assume that the agency's loss function $L^{(A)}_{\Theta_{SOC}}(\cdot,\cdot)$ is such that $L^{(A)}_{\Theta_{SOC}}(a^{(S)}_1, \theta_{SOC})$ increases as $\hat{\theta}^{(S)}$, the optimal estimate of Θ_{SOC} associated with $a^{(S)}_1$, moves away from θ_{SOC} (i.e. the worst is society's estimate of Θ_{SOC} the bigger is the loss for agency and society). As in Section 5, a simple way to define $L^{(A)}_{\Theta_{SOC}}(\cdot,\cdot)$ is the following. Let $L(\cdot,\cdot)$ be a loss function for the agency's estimation problem "Estimate Θ_{SOC}" then define:

$$L^{(A)}_{\Theta_{SOC}}(a, \theta_{SOC}) = \begin{cases} L(\hat{\theta}^{S}, \theta_{SOC}) & \text{if } a = a^{(S)}_1 \\ c_1 & \text{if } a = a^{(S)}_0 \end{cases}. \tag{39}$$

where c_1 is a fixed constant, $c_1 \in [inf\ L(\cdot,\cdot),\ sup\ L(\cdot,\cdot)]$.

In (39) we are implicitly assuming that "society takes no action" is preferred or at least as good as the poorest society's estimate of Θ_{SOC} while society's best estimate of Θ_{SOC} is always preferred to "society takes no action".

Using the agency's loss function in (39) we can define a general class of measures of information loss that interpret information loss as "distance" between society's estimate of Θ_{SOC} and agency's beliefs about the true value of Θ_{SOC}. In particular consider the general formula for disclosure information loss in (32) with:

$$\begin{array}{ll} \diamond & \Theta_{SOC},\ L^{(S)}(\cdot,\cdot) \ \text{ arbitrary} \\ \diamond & h = 1 \\ \diamond & L^{(A)}_{\Theta_{SOC}}(a_1^{(S)}, \theta_{SOC}) \text{ as in (39)} \\ \diamond & \pi_{T^{(S)}}(\cdot) \text{ arbitrary but not degenerate at } sup\, U^{(S)}(\tilde{D}) \end{array} \quad (40)$$

then we have:

$$Information\ Loss = Pr(a_1^{(S)}) \cdot [E_{\Theta_{INT}|D_0}\{L(a_1^{(S)}, \theta_{SOC})\}] + Pr(a_0^{(S)}) \cdot c_1. \quad (41)$$

The information loss in (41) behaves as we expect.

(a) If the released data are such that society has no uncertainty about its estimate of Θ_{SOC} (i.e. $U^{(S)}(\tilde{D}) = 0$) and from the agency's perspective society's estimate is correct with probability 1 ($\pi^{(A)}_{\Theta_{SOC}}(\cdot|D_0)$ degenerate at $\hat{\theta}^{(S)}$), then whatever is the agency's distribution for society's threshold, information loss achieves its minimum ($U^{(S)}(\tilde{D}) = 0$ implies $P(a_1^{(S)}) = 1\ \forall \pi_T^{(S)}(\cdot)$ and we have: information loss$= L(\hat{\theta}^{(S)}, \hat{\theta}^{(S)}) = inf\ L(\cdot,\cdot)$);

(b) If the released data are such that society has no uncertainty about its estimate of Θ_{SOC} (i.e. $U^{(S)}(\tilde{D}) = 0$) and from the agency's perspective society's estimate is very poor with probability 1 ($\pi^{(A)}_{\Theta_{SOC}}(\cdot|D_0)$ degenerate at some value θ^* very far from $\hat{\theta}^{(S)}$), then whatever is the agency's distribution for society's threshold, information loss is very high ($U^{(S)}(\tilde{D}) = 0$ implies $P(a_1^{(S)}) = 1\ \forall \pi_T^{(S)}(\cdot)$ and we have: information loss$= L(\hat{\theta}^{(S)}, \theta^*) \approx sup\, U^{(S)}(\tilde{D})$ for θ^* far enough from $\hat{\theta}^{(S)}$);

(c) If the released data are such that society's desists from estimating Θ_{SOC} ($Pr(a_1^{(I)}) = 0$), then information loss is equal to c_1, intermediate between case (a) and (b).

As in Section 5.2, the term in (41) that measures the consequences for agency and society's of society using its optimal estimate $\hat{\theta}^{(S)}$ can be decomposed as follows:

$$E_{\Theta_{SOC}|D_0}\{L(a_1^{(S)}, \theta_{SOC}) = [U^{(A)} + \Delta] \quad (42)$$

where Δ is given by

$$\Delta = E_{\theta_{SOC}|D_0}[L(\hat{\theta}^{(S)}, \theta_{SOC})] - E_{\theta_{SOC}|D_0}[L(\hat{\theta}^{(A)}, \theta_{SOC})],$$

and $\hat{\theta}^{(A)}$, $U^{(A)}$ are the agency's optimal estimate and the agency's expected loss associated with the agency's optimal estimate of Θ_{SOC} with respect to the loss function $L(\cdot,\cdot)$. As we expect (42) is a increasing function of the agency's uncertainty about the true value of society's target and the "difference" between society's and agency's optimal estimate of Θ_{SOC} as measured by Δ.

If agency and intruder share a quadratic loss function Δ becomes the square of the Euclidean distance between society's and agency's optimal estimates of Θ_{SOC}:

$$\Delta = (\hat{\theta}^{(S)} - \hat{\theta}^{(A)})^2.$$

If we further assume:

$$\diamond\ \pi_{T_j^{(S)}}(\cdot) \sim \ \text{degenerate at } sup\, L(\cdot,\cdot),$$

then we have $P(a_1^{(S)}) = 1\ \forall a_1^{(S)}$ and information loss is a decreasing function of the agency's uncertainty and the square of the Euclidean distance between society's and agency's optimal estimates of Θ_{SOC}:

$$Information\ Loss = [U^{(A)} + (\hat{\theta}^{S)} - \hat{\theta}^{(A)})^2]. \quad (43)$$

The information loss in (43) is closely related with the *society's estimated knowledge* measure of information loss (SEK) proposed by Trottini (2001),

$$SEK = [U^{(A)} + D(\pi^{(S)}_{\Theta_{SOC}}(\cdot|\tilde{D}), \pi^{(A)}_{\Theta_{SOC}}(\cdot|D_0))], \tag{44}$$

where $D(\cdot,\cdot)$ represents a measure of divergence between distributions.
Society's estimated knowledge was proposed by Trottini (2001b) on an ad-hoc basis. The goal was to define a component of a global measure of information loss that could take into account the quality of society's inference and that penalizes bad inferences more than good inferences. It follows from (43) that the SEK belongs to the general class of measures of information loss in (32) if we define the divergence $D(\cdot,\cdot)$ between two probability distributions as the square of the Euclidean distance between the expected value of the two distributions.

10. CONCLUSIONS

In this paper a Bayesian users-agency model is used to define a general class of measures of disclosure risk and information loss. The underlying idea is simple. To assess the effectiveness of a certain masking technique, the agency needs to predict the users' actions when the masked data are release and then quantify the consequences of such actions. The appealing feature of our approach is that it provides a unifying framework to compare and evaluate measures of disclosure risk and data utility of common use in the current research on statistical confidentiality and at the same time it suggests how new and better measures or policy for confidentiality protection can be derived.

Measuring disclosure risk and information loss is a much more difficult task than as is understood today. Our user-agency model shows that the definition of a measure of disclosure risk or information loss requires statistical agencies to make many assumptions about the users' target, the users' prior information, the users' estimation procedure etc. Examples in Section 5 and 9 show that existing measures usually underestimate this uncertainty. For example it is assumed that agency knows exactly the users targets, or the users' threshold for the maximum tolerable expected loss or the users prior distributions, the users' loss functions etc.

We believe that there might be cases where better measures can be derived considering more realistic agency's uncertainties in the model, and cases were assessing disclosure risk and information loss is too difficult, given the multiplicity of users, of users targets and assumptions that agency needs to make. Our intuition is that the latter case is the most common. We expect to explore this issue in some real problem example. If this turns out to be true, we should think to relocate current efforts in the research of statistical confidentiality to develop more and more efficient ways of restricted access versus trying to define new and ad hoc measures of disclosure risk and information loss.

ACKNOWLEDGEMENT

Preparation of this paper was supported by a Marie Curie Fellowship of the European Community program "Improving The Human Research Potential" under the contract number HPMFCT-2000-00463. The content of the paper reflects the author's personal opinion. The European Commission is not responsible for any views or results expressed.

REFERENCES

Blien, U., Wirth, H., and Muller, M. (1992), Disclosure Risk for Microdata Stemming from Official Statistics, *Statistica Neerlandica*, 46, 69-82.

Cassel, C.M. (1976), Probability based disclosure. In T . Dalenius and A. Klevmarken, eds., *Personal Integrity and the Need for Data in the Social Sciences.*

Corsini, V., Franconi, L., Pagliuca, D. and Seri, G. (1998). An application of microaggregation methods to Italian business surveys. *Proceedings of "Statistical data Protection SDP '98"*, 109-113. Eurostat, 25-27 March 1998, Lisbon, Portugal.

DeGroot M.H. (1962), Uncertainty, Information and Sequential experiments, *Annals of Mathematical Statistics*, 3, 404-419

De Waal A. G. and Willenborg L.C.R.J.(2001), Optimal Local Suppression in Microdata, *Journal of Official Statistic, 14*, 421-435.

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. (2001), Risk of statistical confidentiality disclosure: A comparison of masked and synthetic data release, *Research in Official Statistics*, Vol. 3, in press.

Duncan, G.T. and Lambert, D. (1986). Disclosure-limited Data Dissemination. *Journal of the American Statistical Association, 81*, 10-18.

Duncan, G.T. and Pearson, R.W. (1991). Enhancing Access to Microdata While Protecting Confidentiality. *Statistical Science, 6*, 219-239.

Etzioni, R. and Kadane, J.B. (1993). Optimal Experimental Design for Another's Analysis. *Journal of the American Statistical Association, 88*, 1404-1411.

Lambert, D. (1993), Measures of disclosure risk and harm, *Journal of Official Statistics*, Vol. 9, pp. 313-333.

Trottini, M. (2001a), Users' uncertainty, disclosure risk and data utility, Technical Report 02-2001, Departamento de Estadística y Investigación Operativa, Universitat de València.

Trottini, M. (2001b), A decision-theoretic approach to data disclosure problems, *Research in Official Statistics*, (in press)

Willenborg, L. and de Waal, T. (2001), *Elements of Statistical Disclosure Control*, Lectures Notes in Statistics, Vol. 155. Springer Verlag, New York.

APPENDIX (Proofs)

In this Section we proof the results in Sections 5 and 9.

Section 5.1

From the set of assumptions in (12) and the general formula for disclosure risk in (6) we have:

$$P(J=j) = P(J=j|D_0) = \begin{cases} 1 & \text{if } j=1 \\ 0 & \text{if } j \neq 1 \end{cases} ;$$

$$\begin{aligned} \textit{Disclosure Risk} &= Pr(a_1^{(I)}) \cdot E_{\Theta_{INT}|D_0}\{L^{(A)}_{\Theta_{INT}}(a_1^{(I)}, \Theta_{INT})\} \\ &= Pr(a_1^{(I)}) \cdot c \end{aligned} \tag{45}$$

and the agency's disclosure rule is:

Agency's disclosure rule: A data set $\tilde{D}$ is considered for release if and only if

$$\begin{aligned} & Pr(a_1^{(I)}) \cdot c < t_A \leftrightarrow \\ & Pr(a_1^{(I)}) < t_A^{'} \leftrightarrow \\ & \int_{U^{(I)}(\tilde{D})}^{\infty} \pi_T^{(I)}(t)dt < t_A^{'} \leftrightarrow \\ & U^{(I)}(\tilde{D}) > t_A^{''} \end{aligned} \tag{46}$$

which is exactly the disclosure rule proposed by Duncan and Lambert (1986) with $t' = t''_A$.

If in (12) we specify distribution for the intruder's threshold as in (13) then from (6) we have:

$$Pr(a_1^{(I)}) = \begin{cases} 0 & \text{if } U^{(I)}(\tilde{D}) > U^{(I)}(D_{None}) \\ (U^{(I)}(D_{None}) - U^{(I)}(\tilde{D}))U^{(I)}(D_{None}) & \text{if } U^{(I)}(\tilde{D}) \leq U^{(I)}(D_{None}) \end{cases} \tag{47}$$

and taking $c = U^{(I)}(D_{None})$ as in (13), from (45) we obtain the formula for the (intruder's) knowledge gain in (14). Derivation of formula (16) is almost identical and is omitted.

Section 5.1.1

We start with disclosure rule 1. The intruder's target is an indicator function Θ_{INT} such that $\Theta_{INT} = i$ if and only if the target individual belongs to category i. It can be proved (see Duncan and Lambert 1986) that if the prior predictive distribution is exchangeable and the target individual is known to belong to the table the posterior probability that the target individual belongs to category i equals the tabular proportion of category i, i.e.:

$$Pr(\Theta_{INT} = i|\text{released table}) = p_i.$$

Under the loss function $L_1^{(I)}(\cdot,\cdot)$ the intruder's optimal estimate of Θ_{INT} and the intruder's uncertainty are given by:

$$\begin{aligned} \hat{\theta}^I &= argmax_i \; p_i; \\ U^{(I)} &= \gamma(max_i \, p_i) \cdot (1 - max_i \, p_i). \end{aligned}$$

From the set of assumptions in (17) it follows that the agency's disclosure rule is:

Agency's disclosure rule: Release the data if and only if

$$\begin{aligned} &\gamma(max_i \, p_i) \cdot (1 - max_i \, p_i) > t \leftrightarrow (\gamma(\cdot) \text{ is decreasing}) \\ &max_i \, p_i < t_1 \end{aligned}$$

which is exactly disclosure rule 1.

Now disclosure rule 2. Under the loss function $L_2^{(I)}(\cdot,\cdot)$ the intruder's optimal estimate of Θ_{INT} and the intruder's uncertainty are given by:

$$\begin{aligned} \hat{\theta}^I &= argmin_i \; min\{(1-p_i)^2, bp_i\}; \\ U^{(I)} &= min_i \; min\{(1-p_i)^2, bp_i\}. \end{aligned}$$

From the set of assumptions in (17) it follows that the agency's disclosure rule is:

Agency's disclosure rule: Release the data if and only if

$$\begin{aligned} &min_i \; min\{(1-p_i)^2, bp_i\} > t \leftrightarrow \\ &0.5 - (0.25 - t/b)^{1/2} \leq p_i \leq min\,\{0.5 + (0.25 - t/b)^{1/2},\, 1 - t^{1/2}\} \end{aligned}$$

which is equivalent to disclosure rule 2.

Section 5.1.2

We start with Risk 1. The intruder target is to identify a particular respondent I_h in $\tilde{D}$, i.e. Θ_{INT} is an indicator function s.t. $\Theta_{INT} = i$ if and only if R_i belongs to I_h. Under the loss function in (18) the intruder's optimal estimate of Θ_{INT} is the value h' such that $h' = argmax_i \; p_i^h$ and the intruder's uncertainty is $1 - p_{h'}^h$. In other words if the intruder decides to make a link he will link the record I_h with the record in $\tilde{D}$ that has maximum posterior probability of belonging to I_h and the intruder's uncertainty is one minus this maximum probability. Using the agency's distribution for the intruder's threshold in (18) and formula (6) we have:

$$Pr(a_1^{(I)}) = 1 - U^I(tildeD) = p^h_{h'} \ = max_i \ p_i^h$$

and taking $c = 1$ as in (18) from (45) we obtain:

$$DisclosureRisk = max_i \ p_i^h = Risk \ 1,$$

as required.

The expression for Risk 2 follows easily noticing that conditionally on $J = h$ the disclosure risk is equal to Risk 1,

$$DisclosureRisk | J = h = max_i \ p_i^h$$

The assumption that each individual in the released data is equally likely to be the target and formula (6) then imply that

$$Disclosure \ Risk = \sum_{h=1}^{N} N^{-1} \cdot max_i \ p_i^h = Risk \ 2.$$

Now Risk 3. Conditionally on $J = h$ the intruder's loss function is the same than for Risk 1. This imply that conditionally on $J = h$, also the intruder's optimal estimate of Θ_{INT} and the intruder's uncertainty are the same than in Risk 1, i.e.,

$$\begin{aligned} \hat{\theta}_h^{(I)} &= argmax_i \ p_i^h \\ U_h^{(I)} &= 1 - max_i \, p_i^h \end{aligned}$$

If the agency's distribution for the intruder's threshold is degenerate at t^* as in (19) then from (6) we have that:

$$Pr(a_1^{(I)}) = \begin{cases} 0 & \text{if } max_i \, p_i^h < t^{**} \\ 1 & \text{if } max_i \, p_i^h > t^{**} \end{cases} \tag{48}$$

where $t^{**} = 1 - t^*$. In other words if we assume that the agency knows exactly the intruder's threshold for the maximum tolerable risk, then conditionally on $J = h$ the intruder's action from the agency perspective is a degenerate random variable and the conditional disclosure risk takes value 1 if the intruder's uncertainty is less than t^* (i.e if the intruder links I_h with some record in $\tilde{D}$), zero otherwise. From formula (6) the disclosure risk is just the proportion of links that the intruder makes (or in the terminology of Section 5.1.2 the proportion of individuals identified), which is exactly Risk 3.

Section 5.2

Risk of true identification. Conditionally on $J = h$ the intruder's loss function, the intruder's target, and the agency's distribution for the intruder's threshold, are the same than for Risk 3. Therefore conditionally $J = h$, also the intruder's optimal estimate of Θ_{INT}, the intruder's uncertainty and the probability that the intruder uses his optimal estimate are the same than in Risk 3. In particular we have:

$$\hat{\theta}_h^{(I)} = argmax_i \ p_i^h \tag{49}$$

$$Pr(a_1^{(I)}) = \begin{cases} 0 & \text{if } max_i \, p_i^h < t^{**} \\ 1 & \text{if } max_i \, p_i^h > t^{**} \end{cases} \tag{50}$$

where $t^{**} = 1 - t^*$.
The assumption that conditionally on $J = h$: the agency distribution for Θ_{INT} is degenerate at h^* implies that:

$$-E_{\Theta_{INT}|D_0, J=h}\left[L(\hat{\theta}^I, \theta_{INT})\right] = -L(\hat{\theta}^I, h^*) = \begin{cases} 0 & \text{if } \hat{\theta}^I = h^* \\ 1 & \text{otherwise} \end{cases} \tag{51}$$

The conditional disclosure risk given $J = h$ (which is the product of (50) times (51)) is therefore:

$$\mathit{Disclosure\ Risk}\ |J = h = \begin{cases} 1 & \text{if } \hat{\theta}^I = h^* \text{ and } max_i\, p_i^h > t^{**} \\ 0 & \text{otherwise} \end{cases} \tag{52}$$

that can be rewritten as follows:

$$\mathit{Disclosure\ Risk}\ |J = h = \begin{cases} 1 & \text{if } p_{h^{**}} > t^{**} \text{ i.e. if } I_h \text{ is correctly identified} \\ 0 & \text{otherwise} \end{cases} \tag{53}$$

The assumption that each individual in the released data is equally likely to be the target and formula (6) then imply:

$$\begin{aligned} \mathit{Disclosure\ Risk} &= \sum_{h=1}^{N} N^{-1} \cdot [\mathit{Disclosure\ Risk}\ |J = h] \\ &= \%\ \text{respondents correctly identified} = \mathit{Risk\ of\ True\ Identification}. \end{aligned}$$

Section 5.3

Derivation of Risk 4 in Section 5.3 is almost identical to the derivation of Risk 3 in Section 5.2.

Section 9.1

The proof of (33) and (34) is almost identical to the proof of formulas (14) and (16) in section 5.1 and is omitted.

Section 9.1.1

Let p_i be the $M(W)$-vector with all zeros and a one at the i^{th} position and let $I_S = \{i_1, \ldots, i_{M(S)}\}$ be the subset of $\{1, \ldots, M(W)\}$ such that i_j belongs to I_S if and only if w_{i_j} belongs to S.
The assumption that society's prior is uniform on $\{\pi_1(\cdot), \ldots, \pi_{M(W)}(\cdot)\}$ is equivalent to say that

$$Pr(\Theta_{INT} = p) = \begin{cases} 1/M(W) & \text{if } p = p_i,\ i = 1, \ldots, M(W) \\ 0 & \text{otherwise} \end{cases}$$

With the above notation the posterior distribution for Θ_{INT} can be written as:

$$Pr(\Theta_{INT} = p|\text{released table}) = \begin{cases} 1/M(S) & \text{if } p = p_i,\ i \in I_S \\ 0 & \text{otherwise} \end{cases}$$

Under the loss function $L^{(S)}_{\Theta_{SOC}}(e, \theta_{SOC}) = -\sum_{i=1}^{M(W)} log(e_i) \cdot \theta_i$, society's optimal estimate of Θ_{SOC} is the $M(W)$-vector $\hat{\theta}^{(S)} = (\hat{\theta}_1^{(S)}, \ldots \hat{\theta}_{M(W)}^{(S)})$ such that $\hat{\theta}_i^{(S)} = 1/M(S)$ for $i \in I_S$ and $\hat{\theta}_i^{(S)} = 0$ otherwise and society's uncertainty is equal to $log(M(S))$. If all cells but the gran total are suppressed the solution set S coincides with W, the space of tables spanned by $X_1, \ldots, X_N$ and society's uncertainty is $log(M(W)$. From (34) therefore we have

$$\mathit{Information\ Loss} = [log(M(S)) - log(M(W))]/Log(M(W))$$

which is an increasing function of $log(M(S))$ as required.

Section 9.2

$\pi_{T^{(S)}}$ degenerate at $sup\, L_{\Theta_{SOC}}(\cdot,\cdot)$ implies that $P(a_1^{(S)}) = 1\ \forall\, a_1^{(S)}$ and under the set of assumptions in (38) the general formula for information loss in (32) becomes:

$$\begin{aligned} Information\ Loss &= E_{\Theta_{SOC}|D_0}\{L^{(A)}_{\Theta_{SOC}}(a_1^{(S)}, \Theta_{SOC}) = \\ &= E_{\Theta_{SOC}|D_0}\{|\hat{\theta}^{(S)}_{SOC}(\tilde{D}) - \hat{\theta}^{(S)}_{SOC}(D_0)|\} = \\ &= |\hat{\theta}^{(S)}_{SOC}(\tilde{D}) - \hat{\theta}^{(S)}_{SOC}(D_0)| \end{aligned}$$

which is exactly the information loss in (37).

Section 9.3

The proofs for this section are almost identical to the proofs in Section 5.3 and will be omitted.