

A Paradigm for Masking (Camouflaging) Information

Sallie Keller-McNulty, Charles W. Nakhleh¹ and Nozer D. Singpurwalla²

¹*Los Alamos National Laboratories, Los Alamos, NM, USA* ²*The George Washington University, Washington, DC, USA*

Summary

This is an expository paper. Here we propose a decision-theoretic framework for addressing aspects of the confidentiality of information problems in publicly released data. Our basic premise is that the problem needs to be conceptualized by looking at the actions of three agents: a data collector, a legitimate data user, and an intruder. Here we aim to prescribe the actions of the first agent who desires to provide useful information to the second agent, but must protect against possible misuse by the third. The first agent is under the constraint that the released data has to be public to all; this in some societies may not be the case.

A novel aspect of our paper is that all utilities—fundamental to decision making—are in terms of Shannon’s information entropy. Thus what gets released is a distribution whose entropy maximizes the expected utility of the first agent. This means that the distribution that gets released will be different from that which generates the collected data. The discrepancy between the two distributions can be assessed via the Kullback–Leibler cross-entropy function. Our proposed strategy therefore boils down to the notion that it is the information content of the data, not the actual data, that gets masked. Current practice of “statistical disclosure limitation” masks the observed data via transformations or cell suppression. These transformations are guided by balancing what are known as “disclosure risks” and “data utility”. The entropy indexed utility functions we propose are isomorphic to the above two entities. Thus our approach provides a formal link to that which is currently practiced in statistical disclosure limitation.

Key words: Decision-theory; Entropy; Intrusion; Shannon information; Statistical disclosure limitation; Utility.

1 Introduction

The aim of this paper is to propose an architecture for conceptualizing the problem of data confidentiality, also known as “statistical disclosure limitation (SDL)”, and to present an approach for masking information. We distinguish information masking from data masking, because in the latter what is masked are the observed data, whereas in the former what is masked is the information (or knowledge) in the data. The masked information manifests itself in the form of a probability distribution. With that in mind we start by first qualifying the terms masking and information.

1.1 What do Masking and Information Mean?

The word masking signals to us the feature that in the context of data confidentiality, a certain amount of truth has to be revealed, but not the whole truth. A workable meaning of the term information masking will become apparent in Section 5, where we prescribe how masking can be done. Operationally, data masking implies a transformation of the observed data via a location or

scale change or in the case of tabular data, multiplication by a matrix and/or cell suppression; see Dalenius (1977).

The term information is nebulous; often it is synonymous with the term knowledge. To some statisticians like Basu (1975) information is quantified in terms of what it does; also see De Groot (1962). To communication engineers, information is a measure of the amount of effort (or the number of steps) needed to get to the truth [cf. Shannon (1948)]. There are other interpretations of information; see, for example Soofi (1994, 2000), or Cover & Thomas (1991). A notion closely related to information is *entropy*, which has its origins in thermodynamics; it is a measure of the disorder in a system. The paradigm we advocate here makes use of entropy, in the sense that the data that are released have more information entropy than the entropy of the observed data.

We underscore that what is proposed here is a premise for one way to think about the matter of data confidentiality. We do not purport to offer a definitive solution to any specific problem. This would require a deeper appreciation of the nuances associated with the problem and the practical difficulties in implementing any solution that our approach spawns. An overview of the nuances and difficulties in SDL, together with an exhaustive list of references is given by Keller-McNulty & Duncan (2001), which has also served as an inspiration for our work. Whereas we do not offer a definitive solution to a particular problem, what we do propose is an alternative way to conceptualize the problem of data confidentiality and an alternative paradigm for masking, namely, masking information rather than data, which is what is done in SDL.

1.2 Highlights of Previous Work

Much has been written on the topic of SDL, though not in the same vein as that given here; the decision-theoretic paper by Trottni (2001) comes closest in spirit but offers little by way of a working methodology. In Keller-McNulty & Unger (1993) there is a detailed discourse on connecting the conceptual frameworks of statisticians and of computer scientists on matters pertaining to inferential security and database confidentiality. Also therein are references germane to connecting computer science and statistics on the matter of confidentiality. In retrospect, the present paper may be seen as an analogue to the Keller-McNulty & Unger (1993) work, in the sense that here a linkage between communication theory and SDL is explored. Some other notable references that could be of interest to statisticians are the papers by Duncan & Lambert (1986, 1989), Duncan & Pearson (1991), Duncan & Roehrig (2002), and the book by Willenborg & De Waal (2001). The most recent contributions in this topic is Duncan & Stokes (2004), which is also a rich source of the current and previous literature on SDL. Underlying much of the work in many of the above references is the notion of an “*R-U Confidentiality Map*” [cf. Duncan & Fienberg (1999), and Keller-McNulty, Duncan & Stokes (2002)]. It is the balancing of a measure of threats to privacy, called the “*Disclosure Risk R*”, and a measure of the usefulness of the released data, called the “*Data Utility R*” [cf. Lambert (1993)]. A consequence of this balancing act is the principle of “*Disclosure Limitation (DL)*” which strives to mask a database by invoking upon it deterministic and stochastic transformations. Some other approaches to DL include releasing only a sampled subset of the data, including simulated data in the released set, “blurring” the data by grouping and/or adding random error, suppression by excluding certain attributes of the data, swapping by exchanging the values of certain variables between data subjects, and in some cases releasing randomized responses [c.f. Warner (1965)]. This approach is currently popular in the Netherlands, and goes under the acronym PRAM, for *post randomization*; see de Wolf & van Gelder (2004) for an empirical evaluation of PRAM and a list of related references. In effect, the SDL methods involve operations on the observed data as opposed to our approach which involves operations on the probability distributions that generate the data.

2 The Three Agent Set-Up

It is convenient to think of the SDL problem in terms of three agents: a data collection agency, which for us here plays the role of a decision maker \mathcal{D} , a legitimate user of the data \mathcal{S} , and for the lack of a better word, an intruder \mathcal{I} . The decision maker \mathcal{D} has the task of masking the information contained in the data in a manner which reflects a trade-off between \mathcal{S} 's need to know and the risk of \mathcal{I} 's invading the privacy of individuals in \mathcal{D} 's database. Specifically, whereas \mathcal{S} is often interested in properties of the collective that generates the database, \mathcal{I} is interested in connecting specific items with individuals in the database. We assume, as is sometimes the case, that \mathcal{D} is not authorized to reveal the data only to \mathcal{S} but not to \mathcal{I} . That is, the data released by \mathcal{D} is public to all. This assumption may not be true in all societies. The issue that is therefore germane is the amount of information that is acquired by \mathcal{D} and that which is released by \mathcal{D} . If \mathcal{D} reveals all the information that \mathcal{D} possesses, then \mathcal{I} can intrude on individual privacy, and \mathcal{D} 's credibility to uphold confidentiality is tarnished. By contrast, if \mathcal{D} reveals no information, then \mathcal{S} as a provider of credible inputs to policy makers is unable to function. Thus \mathcal{D} is faced with the decision of revealing enough information for \mathcal{S} to function efficiently but not so much that \mathcal{I} is able to intrude. This boils down to balancing risks and rewards as expressed in terms of \mathcal{D} 's utilities. Our goal here is to prescribe a course of actions that \mathcal{D} can take to achieve the said balance.

3 Preliminaries—Entropy and Information

This section, mainly written for those readers unfamiliar with the notions of entropy and information, also serves the purpose of laying out some notation and terminology.

Consider a discrete distribution with support on n points $0, 1, \dots, n-1$. When $n = 1$, the distribution is degenerate with its mass concentrated at a single point, namely 0. Thus there is no uncertainty associated with such a distribution. Alternatively put, this distribution is said to have an *information entropy* of zero.

Suppose now that distribution is uniform over $0, 1, \dots, n-1$. Then the uncertainty associated with such a distribution is a maximum since all the values $0, 1, \dots, n-1$ are equiprobable, with probability $1/n$. The information entropy of this distribution (also known as *Shannon's Entropy*) is $\mathcal{E}_0(n) = -\sum_n (1/n) \log(1/n) = \log n$. In general, the entropy of any discrete distribution having a probability mass p_i at i is defined as $\mathcal{E}(n) = -\sum_i p_i \log p_i$; all logarithms are to the base e . Consequently, the entropy of a distribution having all its mass concentrated at a single point is $1 \log 1 = 0$. Furthermore any discrete distribution having support on n points obeys the inequality $-\sum_i p_i \log p_i \leq \log n$, with equality if and only if the distribution is uniform on the n points. When $n \rightarrow \infty$, the maximum entropy goes to infinity.

Thus to summarize, the entropy of any distribution having support on a discrete set of points, say n , is bounded above by $\log n$, and ranges from 0 (when $n = 1$) to infinity when n is infinite. Figure 1 illustrates the logarithmic behavior of the entropy upper bound.

Observe that for $\mathcal{E}_0(n) = \log n$, the entropy at n^α is, for any $\alpha \in [0, 1]$, $\alpha \log n = \alpha \mathcal{E}_0(n)$; that is, $\mathcal{E}_0(n^\alpha) = \alpha \mathcal{E}_0(n)$. Thus for $\alpha = 0.5$, the entropy with \sqrt{n} is half the entropy with n .

Our proposal is to use the entropy of the discrete uniform distribution as a yardstick for entropy comparisons and information masking. Here the word information means the negative of $\mathcal{E}(n)$; i.e. the *Shannon information* in a discrete distribution having support on n points is $-\mathcal{E}(n)$. Thus for example, the Shannon information for a discrete uniform distribution having support on n points is $-\log n$, the negative of its entropy.

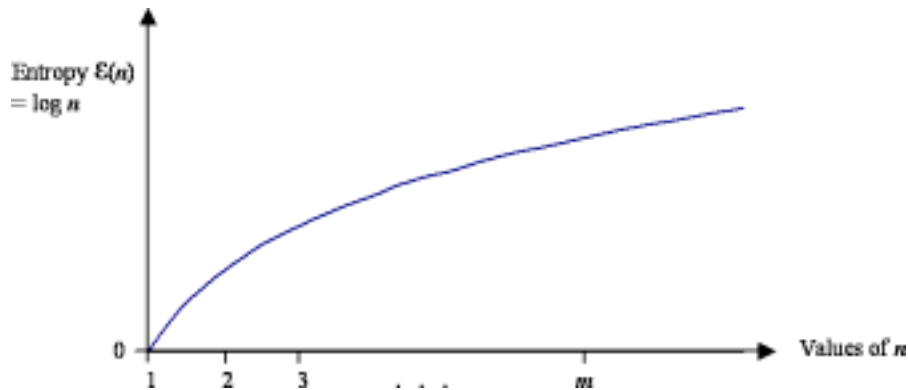


Figure 1. The entropy of a uniform distribution on $[1, n]$.

4 Utilities as Functions of the Entropy

We have stated before, in Section 2, that the masking of information (in the sense described before) entails utilities. Our proposal here is to define utilities as functions of the entropy; this provides an operational meaning to our utilities. This, plus conceptualizing the problem as a three party set-up is the essence of our framework. But how should we calibrate our utilities? In what follows we propose an approach for doing so. But first let us denote by $\mathcal{U}_{\mathcal{D}}(\bullet)$, $\mathcal{U}_{\mathcal{S}}(\bullet)$, and $\mathcal{U}_{\mathcal{I}}(\bullet)$ the utilities of \mathcal{D} , \mathcal{S} and \mathcal{I} respectively. Eventually of course, what matters to us is $\mathcal{U}_{\mathcal{D}}(\bullet)$ since it is \mathcal{D} 's actions that we are attempting to prescribe. However in prescribing $\mathcal{U}_{\mathcal{D}}(\bullet)$, it is necessary that \mathcal{D} be cognizant of $\mathcal{U}_{\mathcal{S}}(\bullet)$ and $\mathcal{U}_{\mathcal{I}}(\bullet)$, because \mathcal{D} 's charter is to assist \mathcal{S} , but to frustrate \mathcal{I} .

Clearly when $\mathcal{E}(n) = \infty$, $\mathcal{U}_{\mathcal{I}}(\mathcal{E}(n))$ is zero, and in principle, $\mathcal{U}_{\mathcal{S}}(\mathcal{E}(n))$ should also be zero. However, in the latter case, if the original data has an infinite entropy to start with, then $\mathcal{U}_{\mathcal{S}}(\mathcal{E}(n))$ may be set to be equal to one (see Section 4.2). With the above in mind, \mathcal{D} should not mask the information in such a way that its entropy is infinite, or else \mathcal{S} will be penalized. Furthermore, when $\mathcal{E}(n) = 0$, $\mathcal{U}_{\mathcal{I}}(\mathcal{E}(n)) = 1$ [and so is $\mathcal{U}_{\mathcal{S}}(\mathcal{E}(n))$], since all the uncertainty has been eliminated and the intruder knows all that is available to \mathcal{D} . Thus $\mathcal{U}_{\mathcal{I}}(\mathcal{E}(n))$ ranges from 1 when $\mathcal{E}(n) = 0$, to 0 when $\mathcal{E}(n) = \infty$. What should $\mathcal{U}_{\mathcal{I}}(\mathcal{E}(n))$ be when $0 < \mathcal{E}(n) < \infty$? This matter is discussed next.

4.1 An Intruder's (Expected) Utility

Our proposal is that $\mathcal{U}_{\mathcal{I}}(\mathcal{E}_0(n)) = (1/n)^\beta$ —which can also be written as $\exp(-\beta\mathcal{E}_0(n))$ —where β , $0 < \beta \leq 1$ is a constant that needs to be specified. The motivation for this choice of the utility function is discussed below.

An intruder is more interested in one (or more) specific individuals in the population, and less so, in general patterns and trends that the population reveals. Thus, intuitively speaking, the smaller the number of individuals, the better is the intruder's ability to pin-point an individual of interest. With $n = 1$, there is no trouble with identification, and thus $\mathcal{U}_{\mathcal{I}}(\mathcal{E}_0(n)) = 1$. When $n = 2$, there are two individuals in the population and the default probability with which \mathcal{I} can identify an individual is

1/2. This probability of identification is the smallest possible, since it is often the case that other auxiliary information will help \mathcal{I} sharpen the identification. Thus we need to increase the crude probability from 1/2 to something larger, and our suggestion is that it be $(1/2)^\beta$, where β is greater than 0 but less than or equal to 1. Values of β close to zero are germane if the amount of auxiliary information available to \mathcal{I} is substantial; this would provide a large boost to 1/2. If \mathcal{I} does not have any auxiliary information then $\beta = 1$; if \mathcal{I} has some (but not much) auxiliary information then β is close to one. If we equate \mathcal{I} 's probability of identification to \mathcal{I} 's utility (as perceived by \mathcal{D}), then $\mathcal{U}_{\mathcal{I}}(\mathcal{E}_0(2)) = (1/2)^\beta$. By a similar line of reasoning with $n = 3$, $\mathcal{U}_{\mathcal{I}}(\mathcal{E}_0(3)) = (1/3)^\beta$ and in general $\mathcal{U}_{\mathcal{I}}(\mathcal{E}_0(n)) = (1/n)^\beta$.

For a nonuniform distribution with entropy $\mathcal{E}(n) < \mathcal{E}_0(n)$, we can easily generalize the utility function using the exponential representation to $\exp(-\beta\mathcal{E}(n))$.

The question now remains as to who is to specify β , $0 < \beta \leq 1$, and on what basis should the β be specified. Our view is that β has to be specified by \mathcal{D} and that it should be based on \mathcal{D} 's appreciation of the amount of auxiliary information available to \mathcal{I} , and also \mathcal{I} 's ability to use such information. This is because \mathcal{I} is not going to reveal his (her) utilities; more important, it is \mathcal{D} 's actions that we are striving to prescribe. As an aside, if \mathcal{D} has reservations about specifying a single value for β , \mathcal{D} may choose to specify a distribution for β , say $\pi(\beta)$, for $0 < \beta \leq 1$. Were \mathcal{D} of the opinion that it is impossible for \mathcal{I} to have auxiliary information whatsoever, then $\beta < 1$, and now $\pi(\beta)$ could be meaningfully well described by a beta density. In what follows we shall refer to β as a *boosting factor*, since the effect of β is to boost upwards the default probability of $(1/n)$.

Thus to summarize, we conceptualize, as a utility function of \mathcal{I} , the relationship

$$\mathcal{U}_{\mathcal{I}}(\mathcal{E}(n)) = \mathcal{U}_{\mathcal{I}}(\log n) = \left(\frac{1}{n}\right)^\beta, \quad 0 < \beta \leq 1,$$

with β specified; see Figure 2 where $\beta = 0.9$.

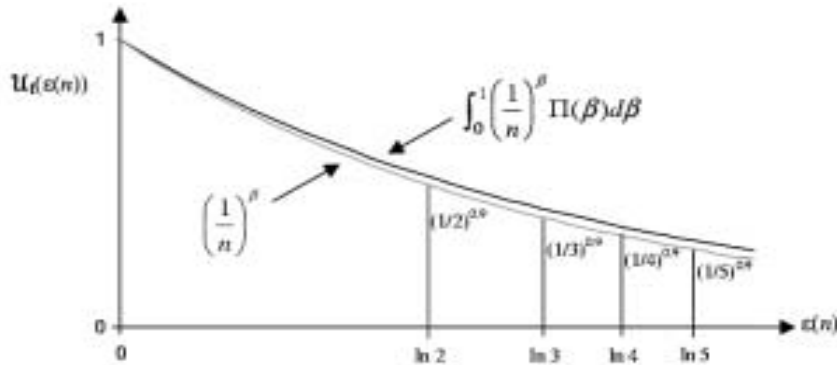


Figure 2. Intruder's (expected) utility versus entropy.

Alternatively, if \mathcal{D} prefers to specify a distribution instead of a specific β , then since utilities are probabilities [cf. Lindley (1985), p. 56, or De Groot (1962), p. 92], we may use the law of total

probability to write

$$U_{\mathcal{I}}(\mathcal{E}(n)) = \int_0^1 U_{\mathcal{I}}(\mathcal{E}(n)|\beta)\pi(\beta)d\beta$$

where $U_{\mathcal{I}}(\mathcal{E}(n)|\beta) = (1/n)^\beta$; thus \mathcal{I} 's (expected) utility, as a function of n , is

$$U_{\mathcal{I}}(\mathcal{E}(n)) = U_{\mathcal{I}}(\log n) = \int_0^1 \left(\frac{1}{n}\right)^\beta \pi(\beta)d\beta.$$

Figure 2 also displays $U_{\mathcal{I}}$ as a function of $\mathcal{E}(n)$ when $\pi(\beta)$ is assumed to be a beta density with a mode of 0.9 and a variance of 0.0107. Our two choices for $U_{\mathcal{I}}(\mathcal{E}(n))$ display a similar pattern.

As an alternative to the above two proposals we may also suppose that $U_{\mathcal{I}}(\mathcal{E}(n)) = \left(\frac{1}{n}\right)^{\frac{n}{n+C}}$, where C is a positive constant. If \mathcal{D} feels that \mathcal{I} has a large amount of auxiliary information, then C should be large; otherwise C is small. With either choice and for sufficiently large n , $\left(\frac{1}{n}\right)^{\frac{n}{n+C}}$ decays faster than $\left(\frac{1}{n}\right)^\beta$ for any fixed β ; this is because $\frac{n}{n+C} \uparrow 1$ as $n \rightarrow \infty$. See Figure 3 for a comparison of the two types of utility functions; here $\beta = 0.2$ and $C = 15$.

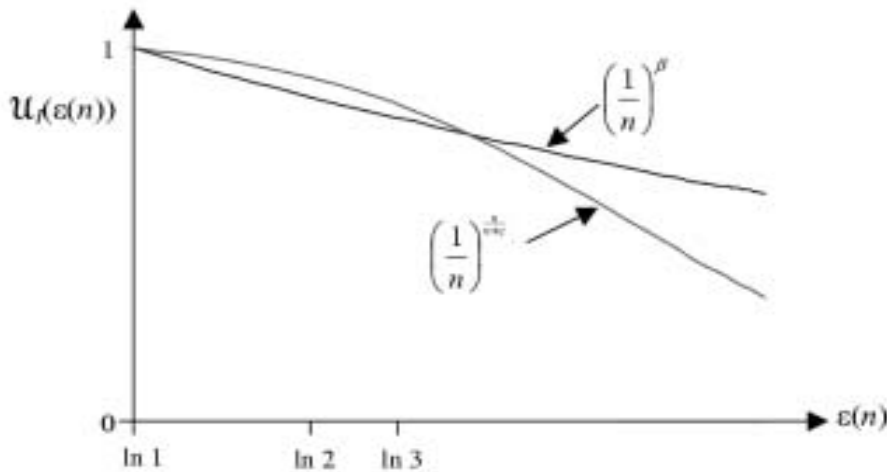


Figure 3. Comparison of two types of utility functions of the intruder.

4.2 A Data User's (Expected) Utility

Suppose that the entropy of the actual data is $\mathcal{E}^*(n)$. Then the utility to \mathcal{S} of receiving this data as is (i.e. without any masking by \mathcal{D}) is the maximum, namely unity. \mathcal{S} will experience a decrease (or loss) in utility whenever the entropy of the released data, say $\mathcal{E}(n)$, deviates from $\mathcal{E}^*(n)$. The further $\mathcal{E}(n)$ is from $\mathcal{E}^*(n)$, the smaller is the utility to \mathcal{S} . There could be several functional forms that are able to capture such a characteristic, one of which is given below; it parallels the utility function that

\mathcal{D} supposes for \mathcal{I} . Specifically

$$U_S(\mathcal{E}(n)|\mathcal{E}^*(n)) = \int_0^1 \exp(-\beta|\mathcal{E}^*(n) - \mathcal{E}(n)|)\pi(\beta)d\beta,$$

where $\pi(\beta)$ has a beta density. In Figure 4, we illustrate the behavior of $U_S(\mathcal{E}(n)|\mathcal{E}^*(n))$, when $\mathcal{E}^*(n) = 3$, and the beta density for β has a mode of 0.1, and a variance of 0.0107. Our choice of 0.1 as a mode for $\pi(\beta)$ reflects a certain symmetric consideration of utilities for \mathcal{I} and \mathcal{S} ; recall that in the case of $U_S(\mathcal{E}_0(n))$, the mode of β was 0.9. Other values for the mode could also be chosen because the symmetry considerations used here are not essential.

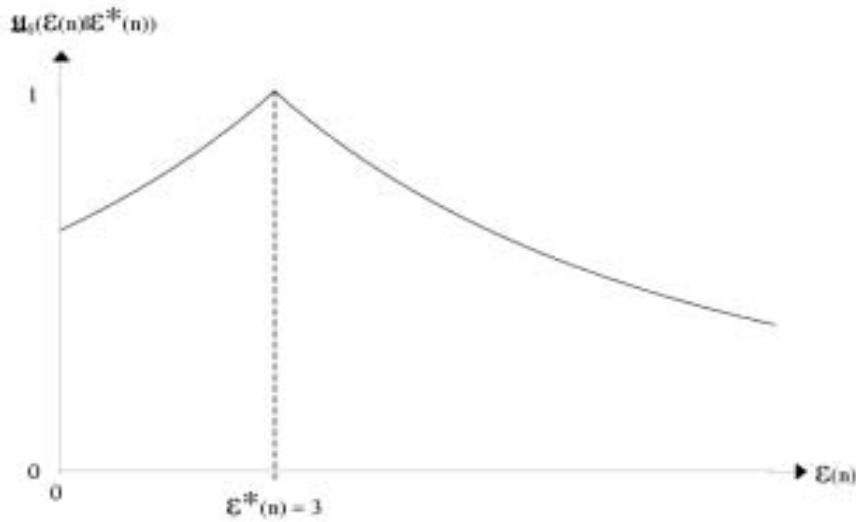


Figure 4. \mathcal{S} 's Utility as a function of the entropy.

The expression for $U_S(\mathcal{E}(n)|\mathcal{E}^*(n))$ is conditioned on a fixed and known value of $\mathcal{E}^*(n)$. Since this value can vary from scenario to scenario, the shape of Figure 4 with a peak at $\mathcal{E}^*(n)$ may not, in the view of \mathcal{D} be representative of the utilities over a broad spectrum of users of statistical data. Thus, it may be more meaningful to assume that $\mathcal{E}^*(n)$ has a distribution with a density $\mathcal{P}(\mathcal{E}^*(n))$ at $\mathcal{E}^*(n) \geq 0$. Averaging out $U_S(\mathcal{E}(n)|\mathcal{E}^*(n))$ with respect to $\mathcal{P}(\mathcal{E}^*(n))$, gives us \mathcal{S} 's (expected) utility as

$$U_S(\mathcal{E}(n)) = \int_0^1 U_S(\mathcal{E}(n)|\mathcal{E}^*(n))d\mathcal{P}(\mathcal{E}^*).$$

A suitable choice for $\mathcal{P}(\mathcal{E}^*(n))$ is a gamma distribution with parameters α (scale) and γ (shape) chosen so that $\mathcal{E}^*(n)$ has a mean γ/α and variance γ/α^2 . Specifically, $\mathcal{P}(\mathcal{E}^*(n))$ has a probability

density function at $\mathcal{E}^*(n)$ of the form

$$\frac{\exp(-\alpha \mathcal{E}^*(n)) (\alpha \mathcal{E}^*(n))^{\gamma-1} \mathcal{E}^*(n)}{\Gamma(\gamma)}, \text{ for } \mathcal{E}^*(n) > 0.$$

Figure 5 illustrates the behavior of $\mathcal{U}_S(\mathcal{E}(n))$ when $\pi(\beta)$ has mode 0.1 and a variance 0.0107, and $\mathcal{P}(\mathcal{E}^*(n))$ has mean 2, and variance 0.1. Note that the effect of $\mathcal{P}(\mathcal{E}^*(n))$ is to smooth out the peak at $\mathcal{E}^*(n)$ in Figure 4. A motivation for the above averaging out will become clearer in Section 4.3, wherein we discuss \mathcal{D} 's utility—the utility that is really germane to masking.

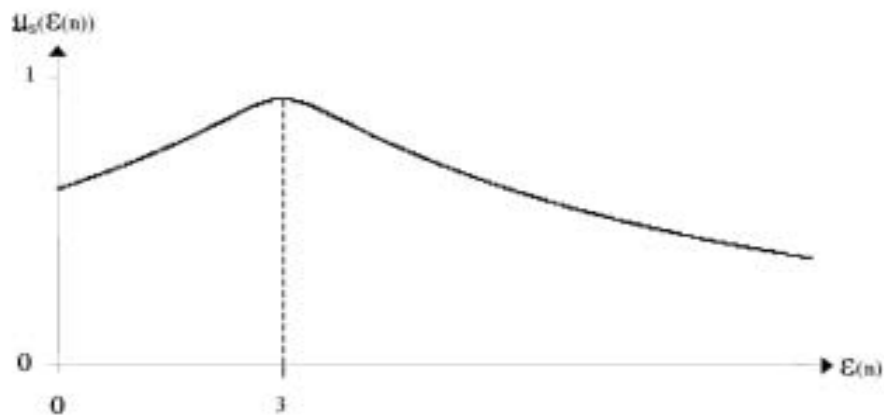


Figure 5. S 's Average utility as a function of the entropy.

4.3 A Decision Maker's Utility

In what follows, we assume that \mathcal{D} is supportive of S 's activities, but is opposed to \mathcal{I} 's actions. Thus \mathcal{I} 's utility is, de facto, \mathcal{D} 's disutility. However, S 's utility should be a part of \mathcal{D} 's utility since S and \mathcal{D} are not adversaries. In view of the above, we may conceptualize \mathcal{D} 's utility as a weighted linear combination of $\mathcal{U}_S(\mathcal{E}(n))$ and $-\mathcal{U}_I(\mathcal{E}(n))$, with weights w_1 and w_2 respectively, where $w_1 + w_2 = 1$. When $w_2 > w_1$, \mathcal{D} chooses to emphasize disutility. This may be meaningful in the sense that greater importance is given to the danger associated with releasing data to \mathcal{I} , than to the benefits received from it via S 's activities. Thus $\mathcal{U}_D(\mathcal{E}(n)) = w_1 \mathcal{U}_S(\mathcal{E}(n)) - w_2 \mathcal{U}_I(\mathcal{E}(n))$, and \mathcal{D} will choose that value of $\mathcal{E}(n)$, say $\hat{\mathcal{E}}(n)$, for which $\mathcal{U}_D(\mathcal{E}(n))$ is a maximum. Figure 6 illustrates the behavior of $\mathcal{U}_D(\mathcal{E}(n))$ with $w_1 = 0.4$, $w_2 = 0.6$, and the $\mathcal{U}_I(\mathcal{E}(n))$ and $\mathcal{U}_S(\mathcal{E}(n))$ are those of Figures 2 and 5, respectively.

The maximum value of \mathcal{D} 's utility occurs when the entropy is $\hat{\mathcal{E}}(n)$; thus $\hat{\mathcal{E}}(n)$ should be \mathcal{D} 's choice for the entropy of the released information. If the actual data has entropy greater than (or equal to) $\hat{\mathcal{E}}(n)$, there is no need for \mathcal{D} to mask it. Otherwise the observed information must be masked to ensure that the entropy of the released information is $\hat{\mathcal{E}}(n)$. The essence of the material of this section is that \mathcal{D} 's utility is driven by \mathcal{D} 's perception of the utilities of \mathcal{I} and S . Also, when the released

information has entropy greater than the observed information, there is a compromise vis-à-vis \mathcal{S} 's ability to produce a trustworthy analysis.

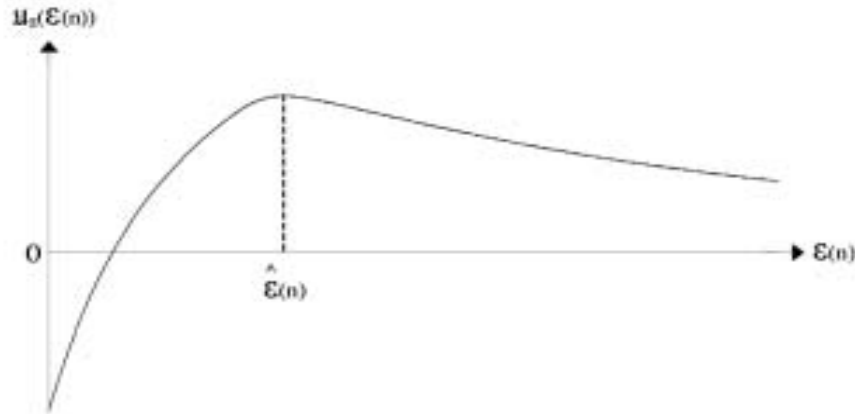


Figure 6. Utility function of the decision maker.

4.4 Isomorphism with the R-U Confidentiality Map: Linkage with SDL

The R-U Confidentiality Map—shown, for example, as Figure 1 of Duncan (2003)—has two ingredients; the disclosure risk R , and the data utility U . As was stated before, in Section 1.2, the former is a measure of the threats to privacy, and the latter a measure of the usefulness to \mathcal{S} of the released data. In our scheme of things, $\mathcal{U}_{\mathcal{S}}(\mathcal{E}(n))$ would encapsulate data utility U , and $\mathcal{U}_{\mathcal{I}}(\mathcal{E}(n))$ the disclosure risk R . Since the underlying concepts driving U and $\mathcal{U}_{\mathcal{S}}(\mathcal{E}(n))$ are similar, we may claim that $\mathcal{U}_{\mathcal{S}}(\mathcal{E}(n))$ is isomorphic to U . Similarly, with R and $\mathcal{U}_{\mathcal{I}}(\mathcal{E}(n))$. Finally, since $\mathcal{U}_{\mathcal{D}}(\mathcal{E}(n))$ is a linear combination of $\mathcal{U}_{\mathcal{I}}(\mathcal{E}(n))$ and $\mathcal{U}_{\mathcal{S}}(\mathcal{E}(n))$, our Figure 6 would be the analogue of the R-U Confidentiality Map, albeit shown on a single scale via $\mathcal{U}_{\mathcal{D}}(\mathcal{E}(n))$. The entropy indexed utilities have an operational interpretation and are easy to appreciate. More important, they are in keeping with the general spirit of balancing U and R that has been advocated in the previous work on data confidentiality.

5 Strategies For Masking Information

Suppose that the observed data is described as the realization of a discrete random variable X , taking values $x = 0, 1, \dots, n$. The x 's need not be restricted to real numbers; they could be intervals. Let $P(X = x) = p(x)$, where $p(x)$ is arrived upon empirically as a relative frequency. Then the observed entropy of the data is

$$\mathcal{E}(n) = - \sum_{x=0}^n p(x) \log p(x).$$

If $\mathcal{E}(n) \geq \widehat{\mathcal{E}}(n)$, where $\widehat{\mathcal{E}}(n)$ is given by the likes of Figure 6, then $p(x)$, $x = 0, 1, \dots, n$ is released, as is, without a mask; there is no gain in utility (to \mathcal{D}) by any form of masking. However, if $\mathcal{E}(n) < \widehat{\mathcal{E}}(n)$, then \mathcal{D} must mask $p(x)$ and release the masked $p(x)$. The question addressed in this section is how should \mathcal{D} mask the observed $p(x)$ as say $\widehat{p}(x)$, where $\widehat{p}(x)$, $x = 0, 1, \dots, n$, encapsulates the information released by \mathcal{D} ? The entropy of $\widehat{p}(x)$ should be $\widehat{\mathcal{E}}(n)$.

We propose here two strategies termed ‘‘Corrupting in a Noiseless Channel’’, and ‘‘Corrupting in a Noisy Channel’’. This terminology resonates with that used in communication and coding theory (see Section 6). The released $\widehat{p}(x)$ could be such that the shape of $p(x)$ may not be preserved. But before describing our strategies for masking observed data it may be instructive to see how one may increase the entropy of some well known discrete distributions. This we do below in Section 5.1.

5.1 Increasing the Entropy of Standard Distributions

Suppose that the data X were to comprise of a discrete distribution. If X is binomial with parameters n and p , denoted $X \sim B(n, p)$ then its entropy can be increased by replacing p with a value nearer to 0.5. For example, if $X \sim B(20, 0.9)$, then its entropy is 1.67; this increases to 2.20 when p is reduced to 0.6.

If X has a Poisson distribution, its entropy is increased by increasing its mean μ . For example with $\mu = 1$ the entropy is 1.30; it increases to 2.20 when $\mu = 5$.

Similarly if X has a geometric distribution, its entropy is increased by decreasing the probability of success p . For $p = 0.4$ the entropy is 1.68; it is 3.25 for $p = 0.1$. The entropy of a geometric distribution with probability of success, p , is given by $\ln\left(\frac{1}{p}\right) + \left(\frac{1-p}{p}\right) \ln\left(\frac{1}{1-p}\right)$. Figure 7 illustrates the behavior of this entropy as a function of p .

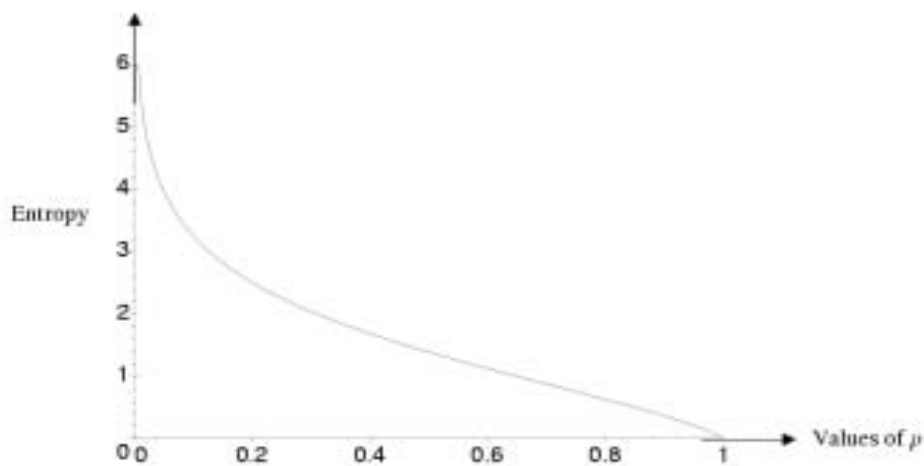


Figure 7. The entropy of a geometric distribution, with parameter p .

If X has a discrete triangular distribution, with mode m , its entropy is increased by shifting the mode towards the center. When X has support $0, 1, \dots, 9$, with a mode at 0, its entropy is 2.08. If

the mode is shifted to 5, the entropy becomes 2.15.

5.2 Corrupting in a Noiseless Channel

There could be several approaches for changing the observed $p(x)$'s in such a way that the resulting entropy is close to $\hat{\mathcal{E}}(n)$, and the released $\hat{p}(x)$'s, $x = 0, 1, \dots, n$ sum to one. We propose below two, both striving to bring the released $\hat{p}(x)$'s closer to $1/n$.

In the first approach, we split the actual $p(x)$'s into two groups, \mathcal{H} and \mathcal{L} where \mathcal{H} contains all the $p(x)$'s greater than $1/n$ and \mathcal{L} contains all the $p(x)$'s that are smaller than or equal to $1/n$. That is,

$$\begin{aligned}\mathcal{H} &= \left\{ p(x) : p(x) > \frac{1}{n} \right\} \\ \mathcal{L} &= \left\{ p(x) : p(x) \leq \frac{1}{n} \right\}.\end{aligned}$$

The corrupted probabilities are then prescribed as:

$$\hat{p}(x) = \begin{cases} (1-u)p(x), & p(x) \in \mathcal{H}, \text{ and} \\ (1+cu)p(x) & p(x) \in \mathcal{L}; \end{cases} \quad (5.1)$$

the constant c ensures that $\sum_{x=0}^n \hat{p}(x) = 1$, and helps the constant u modulate the resulting entropy to $\hat{\mathcal{E}}(n)$. Condition (5.1) above boils down to the requirement that $c = \sum_{\mathcal{H}} p(x) / \sum_{\mathcal{L}} p(x)$; the effect of u is to reduce the $p(x)$'s in \mathcal{H} by a factor of u and to increase the $p(x)$'s in \mathcal{L} by a factor cu . The constant u is determined as a solution to the equation

$$\hat{\mathcal{E}}(n) = - \sum_{x=0}^n \hat{p}(x) \log \hat{p}(x),$$

which now leads us to Equation (5.2), which is the second condition. Specifically,

$$\begin{aligned}\hat{\mathcal{E}}(n) = & -(1-u) \left[\sum_{\mathcal{H}} p(x) \log p(x) + \log(1-u) \sum_{\mathcal{H}} p(x) \right] \\ & -(1+cu) \left[\sum_{\mathcal{L}} p(x) \log p(x) + \log(1+cu) \sum_{\mathcal{L}} p(x) \right].\end{aligned} \quad (5.2)$$

Equation (5.2), which needs to be solved numerically, yields two roots as choices for u . The smaller root is more liable to preserve the shape of the observed probability mass function of X ; the larger root changes the shape. This is obvious from an inspection of the structure of $\hat{p}(x)$; see Equation (5.1). Thus in choosing a root as the value of u , \mathcal{D} can exercise either a shape preserving, or a shape altering strategy when corrupting the $p(x)$'s. The example below illustrates the workings of this approach.

Suppose that X takes values 0, 1, 2 and 3 with respective probabilities 0.5, 0.3, 0.15 and 0.05. The entropy of this distribution is 1.142. Now suppose that we wish to increase this entropy to 1.2. Then, we observe that since $\mathcal{H} = \{0.5, 0.3\}$ and $\mathcal{L} = \{0.15, 0.05\}$, $c = (0.5 + 0.3) / (0.15 + 0.05) = 4$. To find u we need to solve the equation $0.0579 - 1.030u + 0.2 \log(1 + 4u) + 0.8u \log(1 + 4u) + 0.8 \log(1 - u) - 0.8u \log(1 - u) = 0$.

A plot of the left hand side of the above equation is shown in Figure 8. There are two roots, 0.06366 and 0.6289. The smaller root yields 0.468, 0.281, 0.188 and 0.0627 as the corrupted probabilities at 0, 1, 2 and 3 respectively. A comparison with the probabilities of 0.5, 0.3, 0.15 and 0.05 for $X = 0, 1, 2$ and 3 respectively, suggests that $\hat{p}(x)$ with $u = 0.06366$, is shape preserving. By contrast, it can be verified that the root 0.6289 does not preserve the shape of the histogram of X .

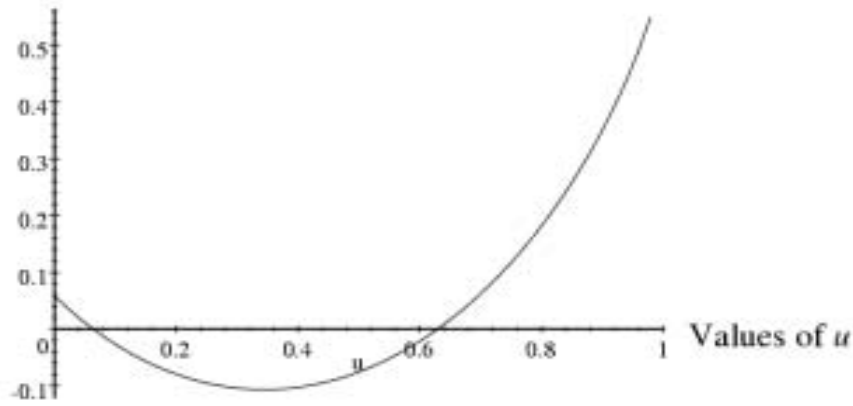


Figure 8. Graph showing solutions of u that yield an entropy of 1.2.

There is another feature to Equation (5.2) that needs mention. Specifically, it can be seen that, as a function of u , $0 \leq u \leq 1$, the right hand side of Equation (5.1) attains a maximum, but that maximum may not be $\log n$, the maximum entropy that any discrete distribution can attain. For example, with regards to the numerical example discussed before, the maximum entropy that the method of Equations (5.1) and (5.2) can yield is 1.31, which is less than $\log 4 = 1.386$, the maximum entropy which a discrete distribution with support over 4 points can achieve. This plus the fact that \mathcal{D} is unable to use the transformation of Equation (5.1) when Equation (5.2) has no roots causes us to seek an alternative method for corrupting $p(x)$. This leads us to the second approach.

With the second approach, we endeavor to reduce the difference between the $p(x)$'s and $1/n$ by some factor u , $0 \leq u \leq 1$. Thus we have

$$\hat{p}(x) = \begin{cases} p(x) - (p(x) - \frac{1}{n})u, & \text{if } p(x) > \frac{1}{n}, \text{ and} \\ p(x) + (p(x) - \frac{1}{n})u, & \text{if } p(x) \leq \frac{1}{n}, \end{cases}$$

or equivalently,

$$\hat{p}(x) = (1 - u) p(x) + \frac{u}{n}, \quad x = 0, 1, \dots, n, \tag{5.3}$$

as our corrupted probabilities. Verify that $\hat{p}(x) = p(x)$ when $u = 0$, and $\hat{p}(x) = \frac{1}{n}$, when $u = 1$. Values of u that are outside the unit interval could result in negative values of $\hat{p}(x)$ and these are inadmissible. Also, $\sum_x \hat{p}(x) = 1$, irrespective of the value of u . Under the transformation of Equation (5.3), the value of u is determined as a solution to the equation

$$\hat{\mathcal{E}}(x) = - \sum_x [p(x) + (\frac{1}{n} - p(x))u] \log [(1 - u) p(x) + \frac{u}{n}]. \tag{5.4}$$

The above equation cannot be further simplified and therefore is cumbersome to solve—particularly so when u is large. However, the corrupting scheme of Equation (5.3) has the advantage that the underlying probabilities can be transformed all the way to the uniform, so that more values of the

entropy are attainable. Furthermore the function

$$\widehat{\mathcal{E}}(x) + \sum_x \left[p(x) + \left(\frac{1}{n} - p(x) \right) u \right] \log \left[(1 - u) p(x) + \frac{u}{n} \right] = 0$$

can be seen to be strictly convex in u , so that there always exists a solution for u , and such a solution is unique (see for example Figure 9).

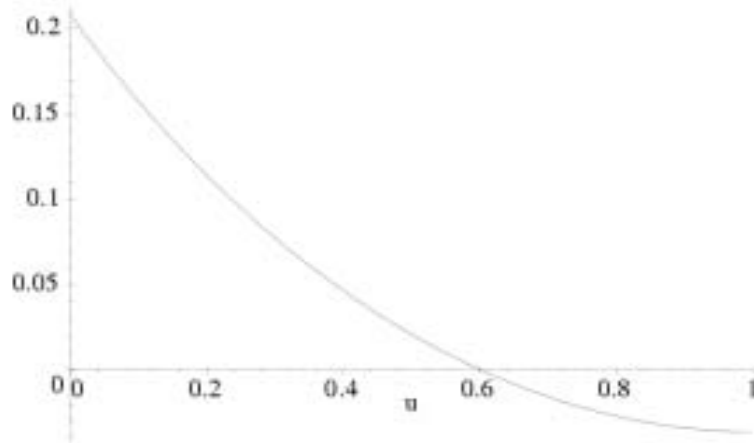


Figure 9. Graph Showing a Solution of u that Yields an Entropy of 1.35.

For the numerical example discussed before, when the method of Equation (5.3) is invoked to obtain an entropy of 1.386 [which the method of Equation (5.1) is unable to provide], u turns out to be 0.6004. This value of u is a solution to Equation (5.4) with $\widehat{\mathcal{E}}(x) = 1.386$, $n = 4$, and the $p(x)$'s being 0.5, 0.3, 0.15, and 0.05, respectively. Figure 9 illustrates the nature of the solution. The corrupted probabilities at 0, 1, 2, and 3 are 0.350, 0.270, 0.210, and 0.170, respectively. Whereas the corrupted values are somewhat shape preserving, they do not do so to the extent that the $u = 0.063$ of the first approach does.

Since the second method is able to yield values of the entropy that cover the range 0 to $\log n$, and since it can also provide a unique value of u for corrupting $p(x)$ —via Equation (5.3)—a question may arise as to why use the first method at all? Our answer lies in the claim that the second method may be harder to implement than the first, because it is more cumbersome to solve Equation (5.4) than it is to solve Equation (5.2). Thus if ease of implementation is a factor, and if \mathcal{D} wishes to exercise the option of either preserving or destroying the shape of $p(x)$, method one is a viable option. However, both approaches entail the numerical solution of equations, namely Equations (5.2) and (5.4). Computer codes for solving these are relatively straightforward to write.

5.3 Corrupting in a Noisy Channel (with Beta Corrupted Noise)

When corrupting in a noiseless channel the $\widehat{p}(x)$'s are predetermined (and thus known to \mathcal{D}), once the $p(x)$'s are at hand. When corrupting in a noisy channel the corrupted probabilities are known only after the noise process has been generated. Thus even \mathcal{D} is unable to know in advance what the

$\hat{p}(x)$'s will be. The scenario here parallels the addition of noise in a coding theory framework (see Figure 10) except that in the latter case it is the message X that gets corrupted by noise whereas here it is the probabilities that get corrupted by noise. Furthermore, in coding theory the corruption is with Gaussian noise whereas here it makes sense to corrupt using noise that has a beta distribution on $(0, 1)$. Let us denote by \mathcal{B} a beta-distributed noise variable. The parameters of this beta distribution can be selected by \mathcal{D} in a manner which preserves or destroys the shape of the distribution of X . A shape preserving corruption would entail choosing the parameters of the beta distribution in such a way that values of the noise process close to zero are emphasized; for destroying the shape, values in the vicinity of one get emphasized.

The procedure for corrupting the $p(x)$'s proceeds as follows.

First we generate n variates from the distribution of \mathcal{B} ; denote these as $b(1), b(2), \dots, b(n)$. We then compute, as an intermediate step, the noise corrupted probabilities

$$p^*(n) = d [b(x)p(x)]$$

where the normalizing factor d ensures that the $p^*(x)$'s sum to one; i.e. $d = (\sum_{x=0}^n b(x)p(x))^{-1}$. Once the $p^*(x)$'s are obtained we use either one of the two approaches of Section 5.2 to obtain final corrupted probabilities $\hat{p}(x)$. The $\hat{p}(x)$'s yield the desired entropy $\hat{\mathcal{E}}(x)$.

The scheme described above is based on the notion of multiplicative noise. An alternative notion is that of additive noise; namely, now

$$p^*(n) = p(x) + b(n) - d$$

where d is given as $d = \frac{1}{n} \sum_{x=0}^n b(n)$.

We are unable to comment as to which of the above two strategies, using a multiplicative or an additive error, is the preferred strategy. Our sense is that it is the former, because of its ability to preserve or to destroy the shape of the distribution of X .

There are, of course, several other strategies by which the $p(x)$'s can be corrupted. One such possibility is based on a variation of the methods of Section 5.2. Specifically, and following the notation of Section 5.2, the corrupted probabilities could be prescribed as:

$$\hat{p}(x) = \begin{cases} (1 - ub(x)) p(x), & p(x) \in \mathcal{H}, \text{ and} \\ (1 + cub(x)) p(x) & p(x) \in \mathcal{L}, \end{cases}$$

with

$$c = \frac{\sum_{\mathcal{H}} b(x)p(x)}{\sum_{\mathcal{L}} b(x)p(x)},$$

or as

$$\hat{p}(x) = (1 - cub(x)) p(x) + \frac{ub(x)}{n},$$

with

$$c = \frac{\sum_{x=0}^n b(x)}{n \sum_{x=0}^n b(x)p(x)}.$$

The $\hat{p}(x)$'s can now be used to solve for u via the two methods of Section 5.2.

6 Connection with Coding Theory: The KL Distance

The masking of information strategies described in Sections 5.2 and 5.3 have a precedence in communications and coding theory (see for example Gallager (1968)). Thus an appreciation of this theory is germane. This section is written with the aim of putting the approaches of Section 5 within a broader perspective. More importantly, the material here also provides a foundation for addressing two other issues. The first pertains to assessing the difference (also known as the Kullback–Leibler (KL) distance or discrimination) between the observed $p(x)$ and the released $\hat{p}(x)$. The second pertains to generalizing the approach of this paper to the bivariate case so that we may be able to mask the Shannon information in tabular data. With regards to the first issue, a question may arise as to the feasibility of developing an information masking procedure that is guided by a desired discrimination between $p(x)$ and $\hat{p}(x)$ via the KL distance. The difficulty with this viewpoint is that in order to arrive upon what the desired discrimination should be, one needs to index utilities by the KL distance. This may not be easy since the KL distance is the expected gain in Shannon information (see Section 6.1), a notion that has an abstract connotation. By contrast our development of entropy indexed utilities is intuitively constructive, and bears a link with that which is currently done in the SDL community.

6.1 Essentials of Coding Theory

In the framework of coding theory, one denotes a message to be sent by X , and the message received by Y . The notions of “message sent” and “message received” are generic. In our particular context, message received can be seen as that random variable whose distribution is $\hat{p}(x)$; and message sent as the random variable X whose distribution is $p(x)$.

The sender (or transmitter) of the message first codes X , as X^* where X^* is some function of X , say $f(X)$ so that $X^* = f(X)$ and then corrupts X^* by a noise factor, say \mathcal{N} ; \mathcal{N} is a random variable, assumed to be independent of X . Thus $Y = g(X^*, \mathcal{N})$ where g is some function of both X^* and \mathcal{N} . The coding of X is de facto a transformation of X by f , and a simple example of this transformation is a matrix mask involving a change in location, the scale, and the shape of X ; i.e. $X^* = A + BX^c$, where the location, scale, and the shape constants, A , B , and C respectively, are specified by the sender of the message. A common strategy for corrupting the coded message is by adding to it Gaussian noise. That is, $Y = X^* + \mathcal{N}$, where \mathcal{N} has a Gaussian distribution with mean 0 and variance σ^2 ; again, σ^2 is specified by the sender of the message. Figure 10 is an illustration of the coding and the corrupting scheme described above. In our case, the purpose of coding and corrupting X is to produce a Y whose entropy is greater than the entropy of X .

In coding theory, the sender of the message needs to inform the receiver of the message, the coding and corrupting schemes; i.e. $A + BX^c$, the receiver of Y needs to be told what f , g , and \mathcal{N} are. Thus, in the case of coding as the receiver needs to be told what A , B and C are, and if corruption is by the addition of Gaussian noise, then the receiver has to be informed as to what σ^2 is. The process of changing an X to a Y is referred to as “passing through a channel disturbed by noise” (see Figure 10). Once Y has been received, the receiver proceeds to reconstruct X by decoding it; because X^* is corrupted by random noise \mathcal{N} , the receiver is unable to precisely decode Y . The decoded message is therefore a good approximation to X in the sense that the entropy of the decoded message is less than the entropy of Y but is close to the entropy of X . The decoding of Y by the receiver of the message entails two actions. One involves deconvolving the received message when \mathcal{N} is additive, and the other involves invoking the *inverse* of f .

The question now arises as to what should f , g , and \mathcal{N} be. Alternatively put, the sender of the message needs to design a communication channel (shown by the dotted box of Figure 10). The sender wants to code and corrupt X so that it is only the intended recipient of the message who can access X via the observed Y through deconvolving and taking an inverse. For this the sender

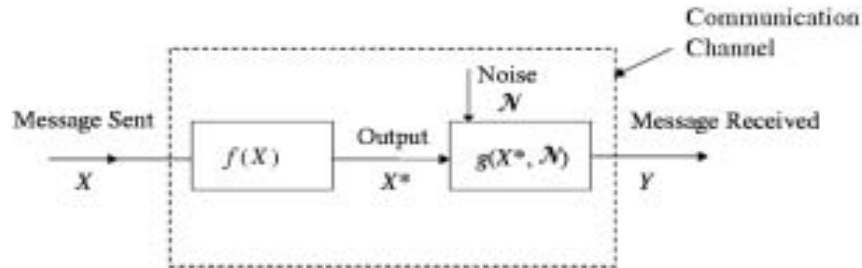


Figure 10. Schemata Showing Coding and Corrupting of Message X .

has to inform the intended receiver as to what f , g , and \mathcal{N} are. The sender has to be cautious in not informing f , g , and \mathcal{N} to the public, for fear that an intruder may also be able to decode Y . The scenario of coding theory has a parallel to, but is unlike that of, data confidentiality wherein \mathcal{D} is mandated to release any masking scheme to all or to none. Bearing this difference in mind, it is useful to gain an appreciation of some underlying ideas behind channel design. We shall see that the ideas of channel design help us address the issues mentioned at the beginning of this Section.

For purposes of discussion let us assume that both X and Y have absolutely continuous distributions with marginal density functions $p(x)$ and $p(y)$, respectively, and a joint density function $p(x, y)$, (assuming that a density exists). For example, the bivariate exponential of Marshall & Olkin (1967) has absolutely continuous marginal distribution functions; however it does not have a joint density function with respect to the two dimensional Lebesgue measure. It is the form of the conditional density function $p(x|y) = p(x, y)/p(y)$, that provides a model for the channel. Specifically, a quantity known as the *mutual information* between X and Y has come to play a prominent role. In particular, $I(X : Y)$ the mutual information between X and Y is defined as

$$I(X : Y) = \mathbf{E}_{p(x,y)} \left[\log \frac{p(x, y)}{p(x)p(y)} \right],$$

where the expectation is with respect to $p(x, y)$. The quantity $I(X : Y)$ has the property that $I(X : Y) \geq 0$ with $I(X : Y) = 0$ if and only if X and Y are independent. The quantity $I(X : Y)$ is Shannon's measure of the expected information about X that is transmitted through a noisy channel. Thus, the minimum value of $I(X : Y)$ corresponds to the case in which a knowledge of Y gives no information about X . A good channel design corresponds to the case in which the mutual information is large. This boils down to finding that $p(x|y)$ for which $I(X : Y)$ is a maximum; the maximum value of $I(X : Y)$ is known as *channel capacity*. Note that finding a $p(x|y)$ corresponds to pinning down the f , g , and \mathcal{N} . It is useful to note that the criterion of maximizing mutual information is equivalent to the principle of maximizing expected utility in decision theory; this equivalence was first noted by Lindley (1956). Thus we see here a parallel between \mathcal{D} 's choosing that entropy which maximizes \mathcal{D} 's utility, and the notion of channel capacity in coding theory.

Before closing this section on some essentials of coding theory it is helpful to make a few other

remarks. The first is the sender's need to corrupt a coded message. Corrupting provides an added security to the message that needs to be transmitted. However, a price has to be paid for this security, because the decoding of Y by the message's receiver does not yield the original X per se; it yields a random variable that is stochastically equivalent to X . The second point to note is that conditioning a random variable X on another random variable Y , decreases the entropy of X [see, for example Cover & Thomas (1991)]. As a consequence, it can be shown that the mutual information is equivalent to the "expected gain in Shannon information", or equivalently, the "expected KL distance" between $p(x|y)$ and $p(x)$ [Kullback & Leibler (1951)]. In our particular context $\hat{p}(x)$ plays the role of $p(x|y)$. Thus the disparity between $p(x)$ and $\hat{p}(x)$ can be assessed via $I(X : Y)$. In the case of tabular data, X and Y can be seen as the two variables whose bivariate distribution generates the data. Using $I(X : Y)$ to develop an information masking scheme is work that remains to be done.

6.2 Masking by Corrupting in Noisy and Noiseless Channels

Whereas \mathcal{D} 's objective is similar to that of the sender of a message in coding theory, \mathcal{D} is handicapped in the sense that \mathcal{D} is not allowed to reveal to \mathcal{S} any coding and corrupting schemes; should \mathcal{D} reveal these to \mathcal{S} , then in our set-up \mathcal{D} is mandated to release them to the public too, which of course would include \mathcal{I} . Thus \mathcal{D} has to strike a balance between the needs of \mathcal{S} and the potential harm that could be done by \mathcal{I} . This \mathcal{D} does by corrupting $p(x)$ in such a way that the entropy of the released $\hat{p}(x)$ is $\hat{\mathcal{E}}(n)$. The general action of coding and corrupting an X to produce a Y with a specified entropy will be known as information masking.

Now it is well known that in the case of a continuous variable X , a change of location does not change the entropy but that a one-to-one transformation of X by scaling does change the entropy. That is, the entropy of a continuous random variable is location invariant but not scale invariant [cf. Cover & Thomas (1991)]. This latter feature is the basis for coding a continuous X . However, as can be easily verified, when X is discrete, any location, scale, or shape transformation of X does not change its entropy. Thus the entropy of a discrete random variable is invariant under any of the usual forms of coding. Consequently, \mathcal{D} is unable to change the entropy by coding alone. This means that corrupting appears to be a viable way by which \mathcal{D} can mask a discrete random variable X . But first let us examine the consequences of corrupting X by a random variable \mathcal{N} .

If the distribution function of \mathcal{N} is absolutely continuous, say a Gaussian, then corrupting X by adding or multiplying it by \mathcal{N} , or even by raising X to a power of \mathcal{N} , will make the masked random variable Y continuous. Whereas these operations will result in the required entropy for Y , the continuity of Y may not be desirable to \mathcal{S} . Thus choosing a discrete random variable \mathcal{N} to corrupt X may be preferable than a continuous one. However, this too has a negative consequence in the sense that the support of Y will be different from the support of X , namely, $0, 1, \dots, n$, and a change of support may also not sit too well with \mathcal{S} 's interests. Thus we need to look at alternative strategies for changing the entropy; one such strategy is the topic of Section 5.

7 Summary and Conclusions

Statistical Disclosure Limitation is an important and a well-discussed problem whose ramifications transcend the original intent under which the subject was conceived, namely, protecting confidentiality in sample surveys. More generally, the topic of SDL can be seen as an element of the problem of *information security*, connecting it with aspects of computer science, communication, coding, and decision theory.

The traditional approaches to SDL have centered around masking the observed data via transformations, simulations, and cell suppressions. These activities have been guided by a balancing of the risks of disclosure against the analytical power provided by the observed data. The balancing of risks

manifests itself via the R - U Confidentiality Map.

In this paper we have proposed a different paradigm for undertaking SDL. Specifically, we propose to mask the information in the data by masking the distribution that generates it. Our paradigm is made workable by the introduction of two features; these constitute the novel aspects of our work. The first is to conceptualize the SDL problem via the actions of three agents \mathcal{D} , \mathcal{S} and \mathcal{I} , with \mathcal{I} being an adversary of \mathcal{D} and \mathcal{S} . In doing so, we have introduced the elements of decision and game theory into the problem. The second feature is to make the first feature operational by defining utilities in terms of Shannon's entropy. The balancing of disclosure risk and analytical power feature of traditional SDL is encapsulated via a maximization of agent \mathcal{D} 's utility function, where this utility function is a linear combination of utility functions of agents \mathcal{S} and \mathcal{I} . Thus what is proposed by us has, in principle, a linkage with what is done in current practice. The paradigm of masking information is also seen from the perspective of coding and communication theory and the relationship between these has been articulated in Section 5.

From an applications point of view, the strategy of how to mask information is guided by the notion that for any given n , the discrete uniform has the maximum entropy. Thus changing the $p(x)$'s in such a way that the resulting distribution gets closer to a uniform achieves the desired goal. With this in mind we have proposed two broad-based approaches, one deterministic that we refer to as corruption in a *noiseless channel* (see Section 5.1), and the other stochastic that we call corruption in a *noisy channel* (see Section 5.2). Each approach spawns two schemes within them, both of which entail some easy to undertake numerical work. One of these schemes offers the user a shape preserving feature; the other does not. Both the noiseless and the noisy approaches result in a masked random variable Y whose support is the same as the original random variable X but whose entropy $\hat{\mathcal{E}}(n)$ is closer to the entropy of a uniformly distributed random variable. Both schemes underlying each approach are not moment preserving. Thus the "analytical power" of the released information is compromised, the extent of the compromise being a function of the Kullback–Leibler distance between the original and the released distribution.

Future work in this arena would entail practical and theoretical matters. Regarding the former, one needs to assess the efficacy of masking distributions over data. The masking schemes are easier to apply and software to do so can be easily developed. The question of how much analytical power is lost with the masking of a distribution remains to be addressed. However, to do so one needs a theoretical framework connecting the Kullback–Leibler distance and loss of analytical power. At a more down to earth level, one needs to develop schemes for masking bivariate and multivariate data. In the bivariate case, the notion of mutual information offers an avenue by which one can formally proceed.

Acknowledgements

The authors are thankful to Asta Manninen, the Editor, and two knowledgeable experts, who served as referees, for their comments which have resulted in the present version of this manuscript. Nozer Singpurwalla's research was supported in part by Grants DAAD 19-01-1-0502 under a MURI and DAAD 19-02-1-0195, The U.S. Army Research Office. He also acknowledges Philip Wilson for discussions and computing help.

References

- Basu, D.V. (1975). Statistical Information and Likelihood. *Sankhyā A*, **37**, 1–71.
 Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory*. New York, NY: Wiley.
 Dalenius, T. (1977). Privacy Transformations for Statistical Information Systems. *Journal of Statistical Planning and Inference*, **60**(309), 73–86.
 De Groot, M.H. (1962). Uncertainty, Information and Sequential Experiments. *Annals of Mathematical Statistics*, **3**, 404–419.
 De Wolf, P.-P. & Van Gelder, I. (2004). An Empirical Evaluation of PRAM. Technical Report, *Statistics Netherlands*, Voorburg.

- The Netherlands.
- Duncan, G.T. (2003). Exploring the Tension Between Privacy and the Social Benefits of Government Databases. Technical Report, School of Public Policy and Management, Carnegie Mellon University.
- Duncan G.T. & Fienberg, S.E. (1999). Obtaining Information While Preserving Privacy: A Markov Perturbation Method for Tabular Data. Eurostat. *Statistical Data Protection '98* Lisbon.
- Duncan, G.T. & Lambert, D. (1986). Disclosure-Limited Data Dissemination. *Journal of the American Statistical Association*, **81**(393), 10–19.
- Duncan, G.T. & Lambert, D. (1989). The Risk of Disclosure for Microdata. *Journal of Business and Economic Statistics*, **7**(2), 207–217.
- Duncan, G.T. & Pearson, R.W. (1991). Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future. *Statistical Science*, **6**(3), 219–239.
- Duncan, G.T. & Roehrig, F. (2002). Mediating the Tension Between Information Privacy. Working Paper.
- Duncan, G.T. & Stokes, S.L. (2004). Disclosure Risk vs. Data Utility : The R-U Confidentiality Map as Applied to Topcoding. *Chance*, **17**(3), 16–20.
- Gallager, R.G. (1968). *Information Theory and Reliable Communications*. New York, NY: Wiley.
- Keller-McNulty, S. & Duncan, G.T. (2001). Disclosure-Limited Statistical Analysis of Confidential Data to Support NSF-Sponsored Digital Government Grant. Los Alamos National Laboratory Technical Report LA-UR-01-1673.
- Keller-McNulty, S.A., Duncan, G.T. & Stokes, S.L. (2002). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Los Alamos National Laboratory Technical Report, LA-UR-01-6428.
- Keller-McNulty, S. & Unger, E.A. (1993). Database Systems: Inferential Security. *Journal of Official Statistics*, **9**(2), 475–499.
- Kullback, S. & Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.
- Lambert, D. (1993). Measure of Disclosure Risk and Harm. *Journal of Official Statistics*, **9**(2), 313–331.
- Lindley, D.V. (1985). *Making Decisions*. 2nd Ed. New York, NY: John Wiley.
- Lindley, D.V. (1956). On the Measure of Information Provided by an Experiment. *Annals of Mathematical Statistics*, **27**, 986–1005.
- Marshall A.W. & Olkin, I. (1967). A Multivariate Exponential Distribution. *Journal of the American Statistical Association*, **62**, 30–44.
- Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 379–423.
- Soofi, E.S. (1994). Capturing the Intangible Concept of Information. *Journal of the American Statistical Association*, **89**(428), 1243–1254.
- Soofi, E.S. (2000). Principle Information Theoretic Approaches. *Journal of the American Statistical Association*, **95**(452), 1349–1353.
- Trottini, M. (2001). Disclosure Risk and Information Loss: A Unifying Approach. Technical Report, Departamento de Estadística e Investigación Operativa, Universitat de Valencia, Spain.
- Warner, S.L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, **60**(309), 63–69.
- Willenborg, L.C. & De Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York, NY: Springer Verlag.

Résumé

Ceci est un article exploratoire. Nous proposons ici un cadre théorique de décision pour traiter d'aspects des problèmes de confidentialité de l'information dans les données diffusées au public. Notre hypothèse de base est que le problème doit être conceptualisé en observant les actions de trois agents: un collecteur de données, un utilisateur légitime d'informations et un intrus. Nous cherchons ici à prescrire les actions du premier agent qui désire fournir des informations utiles au second mais doit se protéger contre une possible mauvaise utilisation par le troisième. La contrainte pour le premier agent est que les données diffusées doivent être entièrement publiques; ce n'est pas forcément le cas dans certaines sociétés.

Un aspect original de l'article est que toutes les utilités—fondamentales pour la prise de décision—sont en terme d'entropie d'informations de Shannon. Aussi ce qui va être diffusé est une distribution dont l'entropie maximise l'utilité attendue du premier agent. Cela signifie que la distribution qui va être diffusée sera différente de ce que génèrent les données collectées. Les divergences entre les deux distributions peuvent être mesurées avec la fonction d'entropie de Kullback–Leibler. Par conséquent la stratégie que nous proposons revient à considérer que c'est le contenu en informations des données, et non les données elles-mêmes, qui reste masqué. La pratique actuelle de "limitation de divulgation statistique" masque les données observées via des transformations ou suppressions de cellules. Ces transformations résultent d'un équilibre entre ce qui est connu comme "risques de divulgation" et "utilité des données". Les fonctions d'utilité indexée d'entropie que nous proposons sont isomorphes des deux entités mentionnées ci-dessus. Aussi notre approche fournit un lien formel avec la pratique courante dans la limitation de divulgation statistique.

[Received March 2004, accepted February 2005]