Entropy as a Measure of Average Loss of Privacy

Luc Longpré[1], Vladik Kreinovich[1], and
Thongchai Dumrongpokaphan[2]
[1]Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
longpre@utep.edu, vladik@utep.edu
[2]Department of Mathematics, College of Science
Chiang Mai University, Chiang Mai, Thailand
tcd43@hotmial.com

Abstract

Privacy means that not everything about a person is known, that we need to ask additional questions to get the full information about the person. It therefore seems to reasonable to gauge the degree of privacy in each situation by the average number of binary ("yes"-"no") questions that we need to ask to determine the full information – which is exactly Shannon's entropy. The problem with this idea is that it is possible, by asking two binary questions – and thus, strictly speaking, getting only two bits of information – to sometimes learn a large amount of information. In this paper, we show that while entropy is not always an adequate measure of the *absolute* loss of privacy, it is a good idea for gauging the *average* loss of privacy. To properly evaluate different privacy-preserving schemes, so also propose to supplement the average privacy loss with the standard deviation of privacy loss – to see how much the actual privacy loss cab deviate from its average value.

1 Formulation of the Problem

Statistical databases: tradeoff between privacy and benefits. Current data mining techniques enable us to extract a lot of useful information from data.

For example, by analyzing information about different medical patients – what were the symptoms, what treatment was applied, what were the results – we can uncover new dependencies and thus, potentially, come up with new recommendations that would lead to a better cure. E.g., in situations where there are two or more possible treatments, by taking into account the patients' age, gender, ethnic origin, habits,, etc., we may be able to describe for which patients which treatment is more promising.

Similarly, by analyzing people's reaction to different movies, books, or foods, researchers have found unexpected correlations that enable them, based on the user's previous selections, to recommend new books (movies, foods, etc.) that will be, with high probability, enjoyed by the user.

However, these benefits come at a price: to be able to achieve them, users need to disclose a large amount of information that they would normally keep private – e.g., details of their illnesses, their vices and habits. This information can be potentially used to harm the person – e.g., insurance companies can use information about a person's health to increase the payments. companies may want to fire people in imperfect health, etc. No matter how we try to anonymize the data, if a database contain enough information about a person, this information can often narrow the person down.

The more detailed information, the larger the benefits – but at the same time the larger the corresponding loss of privacy. Then, the designers and users of large databases must decide how much privacy they are willing to sacrifice to get the corresponding benefits.

To maintain an appropriate tradeoff, we need to be able to gauge privacy and benefits. To be able to formulate the corresponding problem in precise terms, it is necessary to be able to gauge both the loss of privacy and the corresponding gains.

Benefits are the easiest to gauge: simply by asking how much money the user is willing to pay for the corresponding benefit. The monetary equivalents of different benefits have been used in economics, in particular, in economics of medicine and in economics of entertainment.

In contrast, gauging loss of privacy is not easy. Most people have a good understanding how much they are willing to pay to improve their health or to watch a good movie, but they do not have a good feeling for a loss of privacy: in contrast to health and entertainment, the amount of money that people are willing to pay for a certain loss of privacy varies widely.

Since we cannot gauge the loss of privacy based on people's reactions, it is desirable to come up with an objective measure for a loss of privacy.

Entropy as a natural measure of the amount of information. Privacy meas that an outsider is uncertain about the state of the person. Loss of privacy means that this uncertainty decreases – and a complete loss of privacy means that there is no uncertainty left, an outsider knows everything about the given person. Thus, as a measure of privacy, it is reasonable to consider the amount of uncertainty – i.e., the *entropy* of the corresponding distribution; see, e.g., [4, 7].

Entropy can be defined as the average number of binary ("yes"-"no") questions that we need to ask to uniquely determine the alternative. If we have n alternatives, and we do not the probability of each of these alternatives, then the entropy is equal to $S = \log_2(n)$: since after k questions, we have 2^k possible combinations of answers and thus, we can determine 2^k different alternatives.

If we know the probability $p(a_i)$ of different alternatives $a_1, \ldots, a_n$, then the entropy is equal to $S = -\sum\limits_{i=1}^{n} p(a_i) \cdot \log_2(p(a_i))$.

Problem with using entropy to gauge privacy. At first glance, entropy sounds like a reasonable measure of loss of privacy. However, it has a problem; see, e.g., [1, 2]. For example, suppose that a person – e.g., a celebrity – wants to hide her address. We know that she lives in a certain town, on a street where all rich people live, but we do not her house number. The street is long, it has houses numbered from 1 to 2000, with all the numbers used.

The entropy of this situation is $S \approx 11$. In other words, we need 11 binary questions to uniquely determine the celebrity's address.

A user can ask a simple "yes"-"no" question: Is the house number where she lives smaller than 1000 or greater or equal than 1000? Upon receiving the answer, the user get exactly 1 bit of information. This information does not provide the user with much help in finding the desired house: no matter what is the answer, yes or no, the user, instead of a very difficult task of searching through 2,000 possible homes, now has a slightly simpler but still very difficult task of searching through 1,000 possible home. In this case, disclosing one bit of the information did not lead to a big loss of privacy. This makes sense.

But suppose now that the user asks a second question: is the house number smaller than 1001 or greater than or equal to 1001? This is also a one-bit question, and if this was the only question the user asked, it would not bring the user much information.

However, if it so happens that the celebrity lives in the house number 1000, then, by asking these two one-bit questions, the user will learn the celebrity's address: indeed,

- from the answer to the first question, the user will learn that the address is greater than or equal to 1000, and
- from the answer to the second question, the user will learn that it is smaller than 1001,

so 1000 is the only option.

Thus, by asking two simple open-bit questions, each of which does not decrease the privacy much, we can get a serious breach of privacy.

What we show in this paper. In this paper, we show that while entropy may not be a good measure of *exact* loss of privacy, it is a perfect measure of the *average* loss of privacy.

2 Describing the Problem in Precise Terms and the Desired Result

Original situation. In the original situation, we have n possible alternatives $a_1, \ldots, a_n$ describing a person, with probabilities $p(a_1), \ldots, p(a_n)$ of different alternatives. These probabilities should of course add up to 1: $\sum\limits_{i=1}^{n} p(a_i) = 1$.

The amount of privacy in this situation can be described by the entropy

$$S_0 = -\sum_{i=1}^{n} p(a_i) \cdot \log_2(p(a_i)). \tag{1}$$

What is a query and how it decreases privacy. Let us consider a generic query, not necessarily a binary question. Let m denote the number of possible answers to this query. For each alternative a_i, we get one of these m answers.

For every j from 1 to m, let us denote by E_j the set of all the alternatives a_i for which, as a result of this query, we got the j-th answer. The sets $E_1, \ldots, E_m$ form a *partition* of the original set of n alternatives, in the sense that every alternative a_i belongs to one and only one of these sets.

Thus, after receiving the j-th answer, we know that the actual alternative a_i characterizing the person belongs to the set E_j. What is the resulting privacy?

Once we know that the alternative a_i belongs to the set E_j, then the probabilities of all alternatives from outside E_j become zeros, while the probabilities of all alternatives inside E_j change from the original probabilities $p(a_i)$ to new values $p(a_i \,|\, E_j)$. Here, by definition of conditional probability, $p(a_i \,|\, E_j) = \dfrac{p(a_i)}{p(E_j)}$, where

$$p(E_j) = \sum_{k: a_k \in E_j} p(a_k).$$

Thus, in this case, the privacy decreases to the new value

$$S_j = -\sum_{i: a_i \in E_j} p(a_i \,|\, E_j) \cdot \log_2(p(a_i \,|\, E_j)). \tag{3}$$

So, in the case of the j-th answer, the privacy decreases from the original value S_0 to the new value S_j, with a decrease of $S_0 - S_j$. We are interested in the *average* decrease of privacy, i.e., in the average value of this difference

$$\Delta S(\{E_j\}) \stackrel{\text{def}}{=} \sum_{j=1}^{m} p(E_j) \cdot (S_0 - S_j). \tag{4}$$

We want to prove that entropy is a reasonable way of describing the average loss of privacy. In view of the above example, we want to make sure that if we ask two queries, the resulting average loss of privacy cannot exceed the sum of the two average privacy losses corresponding to each of the queries.

In precise terms, we consider two possible queries:

- a query corresponding to a partition $E_1, \ldots, E_m$, and
- a query corresponding to a different partition $E'_1, \ldots, E'_{m'}$.

If we ask both queries, then possible answers to both queries are possible pairs (j, j') of answers to both queries. For each such pair, we know that the alternative belongs to both sets E_j and $E'_{j'}$ and thus, that it belongs to the intersection $E_j \cap E'_{j'}$ of these two sets. So, asking the two queries means that we consider a new partition $\{E_j \cap E'_{j'}\}$ formed by such intersections.

What we want to prove is that the average loss of privacy corresponding to asking both queries does not exceed the sum of average privacy losses corresponding to each of these queries, i.e., that

$$\Delta S(\{E_j \cap E'_{j'}\}) \le \Delta S(\{E_j\}) + \Delta S(\{E'_{j'}\}). \tag{5}$$

Discussion. In the celebrity example, if we assume all 2,000 homes to be equally probable, with probability of each home being the actual celebrity's address equal to 1/2,000, then by asking the corresponding two questions, we can sometimes gain a lot of information. However, the probability of this situation is small (1/2,000), so the average loss of privacy will still be small – on average, it will even less than 2 bits.

3 Proof of Our Main Result

To prove our result, let us find an easier-to-analyze expression for the average privacy loss $\Delta S(\{E_j\})$. This value is computed in terms of entropies S_j corresponding to different possible answers j. For each j, substituting the expression $p(a_i \,|\, E_j) = \dfrac{p(a_i)}{p(E_j)}$ for conditional probability into the formula (3) for S_j, we conclude that

$$S_j = -\sum_{i:a_i \in E_j} \frac{p(a_i)}{p(E_j)} \cdot \log_2\left(\frac{p(a_i)}{p(E_j)}\right). \tag{6}$$

The denominator $p(E_j)$ is a common denominator for all the terms in this sum, so we can simplify the expression by moving this common denominator outside the sum:

$$S_j = -\frac{1}{p(E_j)} \cdot \sum_{i:a_i \in E_j} p(a_i) \cdot \log_2\left(\frac{p(a_i)}{p(E_j)}\right). \tag{7}$$

Logarithm of the ratio is equal to the difference between the logarithms, so we have

$$S_j = -\frac{1}{p(E_j)} \cdot \left(\sum_{i:a_i \in E_j} p(a_i) \cdot \log_2(p(a_i)) - \sum_{i:a_i \in E_j} p(a_i) \cdot \log_2(p(E_j))\right). \tag{8}$$

In the second sum, the term $\log_2(p(E_j))$ does not depend on i and is, thus, a common factor that can be taken out of the sum. The remaining sum is equal to $\sum\limits_{i:a_i\in E_j} p(a_i)$, i.e., equal to $p(E_j)$. Thus, the formula (8) takes the following form:

$$S_j = -\frac{1}{p(E_j)} \cdot \left(\sum_{i:a_i\in E_j} p(a_i) \cdot \log_2(p(a_i)) - p(E_j) \cdot \log_2(p(E_j)) \right). \tag{9}$$

The average loss of privacy is defined as

$$\Delta S(\{E_j\}) = \sum_{j=1}^{m} p(E_j) \cdot (S_0 - S_j). \tag{10}$$

We can separate this sum into a difference of two sums: corresponding to S_0 and corresponding to S_j. Thus, we get

$$\Delta S(\{E_j\}) = \sum_{j=1}^{m} p(E_j) \cdot S_0 - \sum_{j=1}^{m} p(E_j) \cdot S_j. \tag{11}$$

In the first term in the right-hand side, S_0 is a common factor, so this sum takes the form $S_0 \cdot \sum\limits_{j=1}^{m} p(E_j)$. The sum of these probabilities is simply 1, so the first sum in the right-hand side of the formula (11) is simply S_0. Thus, the formula (11) takes the following simplified form:

$$\Delta S(\{E_j\}) = S_0 - \sum_{j=1}^{m} P(E_j) \cdot S_j. \tag{12}$$

Substituting the expression (9) instead of S_j in the formula for the sum, we conclude that

$$\sum_{j=1}^{m} P(E_j) \cdot S_j = -\sum_{j=1}^{m} \sum_{i:a_i\in E_j} p(a_i) \cdot \log_2(p(a_i)) + \sum_{j=1}^{m} p(E_j) \cdot \log_2(p(E_j)). \tag{13}$$

The first sum in the right-hand side of the formula (13) covers all possible alternatives, no matter what answer we got to the query. Thus, this sum is simply equal to $-\sum\limits_{i=1}^{n} p(a_i) \cdot \log_2(p(a_i))$, i.e., to the original entropy S_0. Hence, the formula (13) takes the simplified form

$$\sum_{j=1}^{m} P(E_j) \cdot S_j = S_0 + \sum_{j=1}^{m} p(E_j) \cdot \log_2(p(E_j)). \tag{14}$$

Substituting this expression into the formula (12), we conclude that

$$\Delta S(\{E_j\}) = -\sum_{j=1}^{m} p(E_j) \cdot \log_2(p(E_j)). \tag{15}$$

Thus, after answering the query, the average amount of privacy that we lose is equal to the entropy of the probability distribution of possible answers to the corresponding query.

Let us use this fact to prove the desired property (5). Indeed, according to what we have just found, the left-hand side $\Delta S(\{E_j \cap E'_{j'}\})$ of this inequality is the entropy of the joint distribution of pairs (j, j') of indices. The two terms $\Delta S(\{E_j\})$ and $\Delta S(\{E'_{j'}\})$ are, similarly, the entropies of the corresponding marginal distributions:

- the distribution of the index j corresponding to the first query and
- the distribution of the index j' corresponding to the second query.

It has been proven (see, e.g., [4]) that if we know the two marginal distributions, then the largest possible entropy corresponds to the case when the joint distribution is independent, and the entropy of such independent joint distribution is equal to the sum of the entropies of the original marginal distributions. Thus, for every possible joint distribution, its entropy cannot exceed the sum of the entropies of the two marginal distributions – and this is exactly the desired inequality (5).

The statement is thus proven.

4 Beyond Average Privacy Loss

Need to go beyond the average privacy loss. In general, when we make a decision, we take into account the expected gain or expected loss; see, e.g., [3, 5, 6, 8]. However, it is known that it is also important to take into account risk: there is a difference between earning a dollar and participating in a lottery in which we get nothing or two dollars with equal probability 0.5. To take this difference into account, it is important to consider not just average gain or average loss but also some characteristic describing how different the actual gain or loss can be from the average value.

Idea. In statistics, the most widely used way to gauge this difference is by using the standard deviation σ, which described the mean square deviation from the mean: for a random variable ξ with the mean value μ, standard deviation is defined by the formula $\sigma^2 = E[(\xi - \mu)^2]$, where $E[\cdot]$ denotes the mean value. This formula can also be equivalently written as $\sigma^2 = E[\xi^2] - \mu^2$; see, e.g., [9].

Resulting suggestion. It is therefore reasonable to use a similar characteristic, to gauge not just the mean value of the privacy loss, but also the standard deviation of the privacy loss $S_0 - S_j$.

The standard deviation the difference between the constant S_0 and the resulting privacy S_j is simply equal to the standard deviation of the privacy values, i.e., to the value σ for which

$$\sigma^2 = \sum_{j=1}^{m} p(E_j) \cdot S_j^2 - (\overline{S})^2, \tag{16}$$

where

$$\overline{S} = \sum_{j=1}^{m} p(E_j) \cdot S_j = S_0 + \sum_{j=1}^{m} P(E_j) \cdot \log_2(E_j).$$

Acknowledgments

This work is supported by Chiang Mai University, Thailand. This work was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

References

[1] M. Ceberio, G. Xiang, L. Longpré, V. Kreinovich, H. T. Nguyen, and D. Berleant, "Two Etudes on Combining Probabilistic and Interval Uncertainty: Processing Correlations and Measuring Loss of Privacy", *Proceedings of the 7th International Conference on Intelligent Technologies InTech'06*, Taipei, Taiwan, December 13–15, 2006, pp. 8–17.

[2] V. Chirayath, L. Longpré, and V. Kreinovich, "Measuring privacy loss in statistical databases", In: H. Leung and G. Pighizzini (Eds.), *Proceedings of the Workshop on Descriptional Complexity of Formal Systems DCFS 2006*, Las Cruces, New Mexico, June 21–23, 2006, pp. 16–25.

[3] P. C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.

[4] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[5] R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

[6] H. T. Nguyen, O. Kosheleva, and V. Kreinovich, "Decision making beyond Arrows 'impossibility theorem', with the analysis of effects of collusion and mutual attraction", *International Journal of Intelligent Systems*, 2009, Vol. 24, No. 1, pp. 27–47.

[7] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.

[8] H. Raiffa, *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.

[9] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2011.