Evidence Based Anonymization

Nobuaki HOSHINO*

July 27, 2015

Abstract

No individual shall be identified, which is a typical legal condition for disseminating microdata. This unidentifiability, however, has not been clearly defined in a technical sense. Therefore the present paper technically states a method to decide whether given data are identifiable or not by measuring disclosure risk. The existing theory of disclosure risk lacks the method of deciding its critical value; the present article statistically estimates it by observing that a society has not recognized successful identification. Our method employs observable facts as an evidence of anonymity.

Key words: Population unique, Privacy, Statistical Disclosure Control.

1 Introduction

In many countries no individual is legally allowed to be identified when a statistical agency publishes microdata. For example, in Japan, Statistics Law defines Anonymized Data, which is a scientific use file, so that no individual shall be identified. Hence data publishers need to decide whether a given data set is identifiable or not.

This decision has been made in a rather subjective manner. Technical efforts to formalize this decision result in many measures of re-identification risk, but no objective method to decide a threshold of those measures seems known. This is so because those researches let this decision depend on a personal preference under a tradeoff between re-identification risk and utility of data; see e.g. Duncan et al. (2001). Consequently, data publishers subjectively assess identifiability.

However, identifiability does not depend on a personal preference. When data are identifiable, they are so regardless of the preference of their publisher. Also laws say nothing about the utility of data. In other words, no identifiable data set is permitted to be disseminated even if they are highly useful. Therefore in practice the true goal of anonymization is not to take the balance of risk and utility but to maximize utility within an acceptable range of risk. Hence the decision of a threshold of a risk measure is a primary issue, which needs to be objectively dealt with.

The present article proposes to decide such a threshold by observing whether published data have been identified or not. These observations are statistical samples on idenfiability, and they carry information on a threshold of a risk measure under an appropriate statistical model. Hence after constructing a statistical model of identification, we statistically estimate the threshold of acceptable risk.

*School of Economics, Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan. E-mail: hoshino@kenroku.kanazawa-u.ac.jp

2 Statistical model of identification

This main section describes our method to decide whether an individual is idenfiable or not. The first subsection technically explains the meaning of unidentifiability. The second subsection presents a model in which uncertainty on identification is expressed by a parameter. The third subsection statistically estimates this paramter. The fourth subsection clarifies our methodology to select key variables (quasi-identifiers), because our measure of disclosure risk heavily depends on the selection of key variables.

2.1 Definition of identifiability

An effort on modeling (re-)identification can be seen in Marsh et al. (1991). They argue that the probability of identification is the product of the following probabilities:

$$\Pr(\text{actual identification}) = \Pr(\text{success of identification}|\text{trial of identification}) \Pr(\text{trial of identification}). \tag{1}$$

In eq. (1), the event of "actual identification" is regarded as the joint event of "success of identification" and "trial of identification". This discrimination between "actual identification" and "success of identification" corresponds to different legal concepts of anonymity or unidentifiability.

Absolute anonymity, using a German legal term, is a state where the possibility of identification is eliminated with no doubt. We regard this state as equivalent to a state that

$$\Pr(\text{success of identification}|\text{trial of identification}) = 0. \tag{2}$$

De facto anonymity, which is also a German legal term, is a state where the cost of identification dominates the benefit of identification. In this case the probability of the trial of identification should be low, and thus we regard this state as equivalent to a state that Pr(actual identification) is low.

The present article focuses upon the assessment of the absolute anonymity. That is, we evaluate whether the conditional probability of "success of identification" given "trial of identification" is zero or not.

To evaluate this Pr(success of identification|trial of identification), Marsh et al. (1991) propose the following factorization:

$$\Pr(\text{success of identification}|\text{trial of identification}) = \Pr(a) \Pr(b|a) \Pr(c|a,b) \Pr(d|a,b,c), \tag{3}$$

where the events from a to d are

(a) Information possessed by an attacker (who tries to identify an individual) has the same quality as that of a published file.

(b) A published file contains an individual.

(c) An individual is a population unique.

(d) A population unique is verified to be so.

If we can evaluate the probabilities of the right hand side of eq. (3), we can obtain the conditional probability of "success of identification" given "trial of identification".

However, no one can plausibly evaluate $\Pr(d|a,b,c)$ in general. This is because we can not know unobservable abilities of latent attackers in a society. Elliot et al. (2010, 2011) claim a thorough investigation of such abilities, but uncertainty must remain.

Now let us be reminded that we just would like to know whether eq. (2) holds or not. This evaluation is far easier than to evaluate the conditional probability of "success of identification" given "trial of identification".

Hence we rewrite eq. (3) as

$$\Pr(\text{success of identification}|\text{trial of identification}) = \Pr(a,b,c)\Pr(d|a,b,c). \tag{4}$$

Then we can see that eq. (2) holds if and only if at least one of $\Pr(a,b,c)$ and $\Pr(d|a,b,c)$ is zero. On data for scientific purposes, $\Pr(a,b,c)$ is usually positive. Consequently our usual assessment on unidentifiability reduces to a decision whether $\Pr(d|a,b,c)$ equals zero or not. Since the direct evaluation of $\Pr(d|a,b,c)$ is hopeless, we will estimate whether $\Pr(d|a,b,c)$ equals zero or not.

2.2 Model for discerning identifiability

Based on our argument so far, we would like to discern whether

$$\Pr(d|a,b,c) = 0 \tag{5}$$

or not, since eq. (5) is sufficient for an unidentifiable state.

To do so, we note that $\Pr(d|a,b,c)$ depends on the event of (a,b,c), and $\Pr(a,b,c)$ can be evaluated as we will see. The increment of $\Pr(a,b,c)$ implies that more information about population uniques is published. The more information exists, more easier the verification of population uniqueness should become. Hence the conditional probability of d given (a,b,c) should be monotonically increasing as $\Pr(a,b,c)$ increases. If so, there exists nonnegative β such that

$$\Pr(a,b,c) \leq \beta \Leftrightarrow \Pr(d|a,b,c) = 0. \tag{6}$$

Then we conclude that the assessment of identifiability reduces to the evaluation of $\Pr(a,b,c)$, since eq. (2) is tantamount to $\Pr(a,b,c) = 0$ or $\Pr(d|a,b,c) = 0$.

In the model (6), $\Pr(a,b,c)$ can be interpreted as the easiness of identification. This is a type of re-identification risk measure, and its threhold β is unknown. We decide it by statistical estimation in the following.

2.3 Observational model of identification

For the statistical estimation of β in eq. (6), we need an observation which carries information on β. Hence we would like to observe the event of d or the success of identification. However, it may not always be recognized by a society; a successful attacker may hide. Therefore we discriminate "actual identification" from its social recognition. Let a random variable X be 1 when "actual identification" is socially recognized, and 0 otherwise. That is,

$$\Pr(X=1) = \Pr(\text{recognized}|\text{actual identification})\Pr(\text{actual identification}).$$

Then from eq. (1) and eq. (3),

$$\begin{aligned}\Pr(X=1) &= \Pr(\text{recognized}|\text{actual identification}) \\ &\quad \times \Pr(a,b,c,d)\Pr(\text{trial of identification}). \end{aligned} \tag{7}$$

Further, let us write the evaluated value of $\Pr(a,b,c)$ as γ, and write

$$p(\gamma) = \gamma \Pr(d|a,b,c)\Pr(\text{recognized}|\text{actual identification})\Pr(\text{trial of identification}). \tag{8}$$

Then

$$\Pr(X=1) = \begin{cases} p(\gamma) & \text{if } \gamma > \beta \\ 0 & \text{if } \gamma \leq \beta. \end{cases} \tag{9}$$

If $p(\gamma)$ is positive, the observed value of X carries information on β, and we can estimate β from X's. Actually $p(\gamma)$ is positive when both

$$\Pr(\text{recognized}|\text{actual identification})) > 0 \tag{10}$$

and

$$\Pr(\text{trial of identification}) > 0 \tag{11}$$

hold. The first condition (10) should be satisfied because an attacker has an incentive to show off their success of identification. Also hiding through is not always possible. The second condition (11) should also be satisfied because of an incentive to do so. Hence we regard that $p(\gamma)$ is positive. It is worth mentioning that we assume no specific form of $p(\cdot)$.

Suppose that there are n past experiences of publishing confidentialized data. We regard these as independent samples from the model (9). For the i-th, $i = 1, 2, \ldots, n$, sample we measure $\Pr(a,b,c) = \gamma_i$ and observe the social recognition of actual identification $X_i = x_i$. Write the likelihood of the observations as $\ell(\beta)$. To simplify our argument we assume that $\gamma_1 > \gamma_2 > \cdots > \gamma_n$.

Now we consider the maximum likelihood estimator $\hat{\beta}$ of the threshold. If there exists an integer m such that $x_{m-1} = 1, x_m = x_{m+1} = \cdots = x_n = 0$, then $\ell(\beta) = 0$ for $\beta \geq \gamma_{m-1}$, $\ell(\beta) \propto p(\gamma_{m-1})$ for $\gamma_{m-1} > \beta \geq \gamma_m$, and $\ell(\beta) \propto p(\gamma_{m-1}) \prod_{j=m}^{i}(1 - p(\gamma_j))$ for $\gamma_i > \beta \geq \gamma_{i+1}, i \geq m$. Hence $\gamma_{m-1} > \hat{\beta} \geq \gamma_m$ because $p(\gamma)$ is positive. If there exists no social recognition of actual identification, then $\hat{\beta} \geq \gamma_1$.

In general we denote the lowest easiness of identification among samples with social recognition of actual identification by γ^-. If there has been no such recognition, let γ^- be 1. Also among samples with the easiness that is lower than γ^- we denote the highest easiness of identification by γ^+. Then $\gamma^+ \leq \hat{\beta} < \gamma^-$.

2.4 How to select key variables

Now we would like to establish the method of evaluating $\Pr(a,b,c)$. However, beause of space restriction, we only discuss the selection of key variables, on which the number of population uniques depend. Readers who are interested in the estimation of population uniques should refer to Hoshino (2001, 2009).

Since the maximum likelihood estimation relatively compares $\Pr(a,b,c)$, we have to select key variables with the same way among observations. However, to do so is not a simple task.

For example, Elliot et al. (2010, 2011) claim a comprehensive survey of information about individuals in a society to formally select key variables. The author never denies the value of such information, but their argument does not directly result in the best selection of key variables in any sense. Fung et al. (2010) describe the selection of key variables as "an open problem".

The present article selects key variables to best estimate β. Existing researches can not optimize the selection because they do not consider the aftermath of evaluating population uniques.

Suppose that there are k variables in a published file. Then there are 2^k ways to select key variables in theory. The number of population uniques can be evaluated in each way, and we write the order statistics of these numbers as $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(2^k)}$. Then the issue of the selection of key variables is nothing but the selection of a rank among $(1, 2, \ldots, 2^k)$, over which attackers are distributed subject to their information about individuals.

For given data we evaluate $\Pr(a, b, c)$ at a selected rank r, and compare it with γ^+, which is also evaluated with the same way of the selection of key variables. Suppose that $\Pr(a, b, c)$ at the r-th rank is smaller than γ^+. Then, for fixed $\Pr(a, b)$, the given data should be safe against attackers who lie on ranks smaller than r, since $u_{(i)} \leq u_{(r)}$ for $i \leq r$. The given data, nevertheless, have no evidence of safety against attackers who lie on ranks larger than r.

Thus one might think that we should select the largest rank: 2^k. However, an attacker may not exist on the 2^k-th rank. If so, an observed X of eq. (9) carries no information on the safety of $\Pr(a, b, c)$ at the 2^k-th. Hence, considering the distribution of attackers over the ranks, we should select the largest rank on which an attacker exists.

Theoretically the best way to select key variables has been described, but in practice, the distribution of attackers is unknown. Therefore we have to estimate the maximum of the distribution of attackers. The precise estimation of a maximum is, however, known to be difficullt, and an errorneous estimate of the maximum rank leads to unstable $\hat{\beta}$. Hence, as a second best way, we should estimate a percentile, which is less difficult. For example to estimate the 99th percentile of the distribution of attackers should be practical as in the case of financial risk: Value at Risk (VaR).

Of course the quantitative evaluation of such a percentile is virtually impossible since we can not know latent attackers. Hence in practice we select key variables whose information is publicly known; this policy implies that a large percentile is selected.

3 Concluding remarks

The implication of our argument in Section 2 is clear: A given data set is publishable if its $\Pr(a, b, c)$ does not exceed γ^+, since it has a statistical evidence of unidentifiability.

For calculating γ^+, it is essentially important to select past examples that share the same threshold β. The threshold changes when an element of identification that is not measured by $\Pr(a, b, c)$ changes. Such elements include the amount of people whose information is publicly known. To control the change of unmeasured elements, we have to separate them in the estimation of β. For examle, Japanese households' case can not be an evidence of American enterprises'; very different populations should have different β. What if there has been no past example that can be an evidence? Then begin with publishing apparently safe data; a clinical trial decides the threshold of some dose by gradually increasing risk.

The key idea of the present article is to employ a statistical model of identification that is based on observable facts. This idea is applicable to decide a threshold of another risk measure, which should be appealing.

The present article has described one method to objectively decide whether given data are identifiable or not. Subjective decision may employ past experiences implicitly. On the contrary our method explicitly employs past experiences as a statistical evidence. In this sense the proposed method should be called Evidence Based Anonymization.

Acknowledgements

The argument of the present article is an extract from the author's discussion paper written in Japanese. This research has been supported by Kakenhi grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

[1] Duncan, G., Keller-McNulty, S.A. and Stokes, S.L. (2001) Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 121, National Institute of Statistical Sciences, Durham, North Carolina.

[2] Elliot, M., Lomax, S., Mackey, E. and Purdam, K. (2010) Data Environment Analysis and the Key Variable Mapping System. *Privacy in Statistical Databases*, Domingo-Ferrer, J. and Magkos, E. (Eds.), LNCS 6344, 138–147, Springer-Verlag, Berlin Heidelberg.

[3] Elliot, M., Mackey, E. and Purdam, K. (2011) Formalizing the Selection of Key Variables in Disclosure Risk. *Int. Statistical Inst.: Proceedings of the 58th World Statistical Congress*, 2777–2784.

[4] Fung, B.C.M., Wang, K., Fu, A.W.C and Yu, P.S. (2010) *Introduction to Privacy-Preserving Data Publishing*, CRC Press, New York.

[5] Hoshino, N. (2001) Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment, *Journal of Official Statistics*, **17**, 499–520.

[6] Hoshino, N. (2009) The Quasi-multinomial Distribution as a Tool for Disclosure Risk Assessment, *Journal of Official Statistics*, **25**, 269–291.

[7] Marsh, C., Skinner, C., Arber, S., Penhale, P., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The Case for a Sample of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society*, Series A, **154**, 305–340.