JJSDS manuscript No.
(will be inserted by the editor)

A Firm Foundation for Statistical Disclosure Control

Nobuaki Hoshino

Received: date / Accepted: date

Abstract The present article reviews the theory of data privacy and confidentiality in statistics and computer science, in order to modernize the theory of anonymization. This effort results in the mathematical definitions of identity disclosure and attribute disclosure applicable to even synthetic data. Also differential privacy is clarified as a method to bound the accuracy of population inference. This bound is derived by the Hammersley-Chapman-Robbins inequality, and it leads to the intuitive selection of the privacy budget ϵ of differential privacy.

Keywords: Differential Privacy, Population Unique, Privacy Budget, Synthetic Data

1 Introduction

The current practice of publishing official statistics faces distrust about the protection of identity. President's Council of Advisors on Science and Technology (2014, pp. 38-39) states that "anonymization of a data record might seem easy to implement," but "as the size and diversity of available data grows, the likelihood of being able to re-identify individuals grows substantially," and "(anonymization) is not robust against near-term future re-identification methods."

The background of these statements seems the realized failures of anonymization in a private sector. Anonymization is not easy to implement at all. It actually requires artisanship for a future-proof data product. Therefore, apprentices have made errors.

Statisticians should recognize that more efforts to theorize the artisanship of anonymization. So far the statistical theory of anonymization lacks the firm definition of anonymity, which should also cause the distrust.

In computer science, Dwork (2006) proposes the notion of Differential Privacy (DP), which is the package of a clear definition of data protection and a method

Nobuaki Hoshino
School of Economics, Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan.
Tel.: +81-76-264-5421
Fax: +81-76-264-5444
E-mail: hoshino@kenroku.kanazawa-u.ac.jp

easy to implement. These two factors are what statisticians' artisanship lacks. Accordingly, researches on DP have exploded. Zhu et al. (2017) guide us around a part of them.

Even the practice of official statistics has been affected by DP. Reiter (2019) favorably introduces the role that DP can and does play in official statistics. However, as Ruggles et al. (2019) state, DP "goes far beyond what is necessary to keep data safe under census law and precedent."

The goal of DP is more stringent than that of traditional statistical practices. Hence DP tends to result in useless data for scientific purposes. Ruggles and others, in particular Bambauer et al. (2013), criticize this little concern about data users. Weak statistical protection combined with other institutional measures constitutes our wisdom to publish usable data.

However, Dwork's easy protection method just adds a noise from the Laplace distribution; Kotz et al. (2001) provide a monograph on this distribution. Additive noise has been a part of statistical anonymization methods. Also DP is defined using the likelihood ratio, with which statisticians are familiar. We may be able to modernize the statistical theory of anonymization with the help of DP.

Therefore, the present article reviews the theory of data protection in statistics and computer science. This effort results in the mathematical definition of the traditional targets of statistical protection: Identity disclosure and attribute disclosure. The scope of these new definitions embraces even synthetic data. Also the key challenge of DP, the selection of the level of protection, is solved using statistical theories. These contributions develop after we realize that the disclosure of information about a population unit is not necessarily confined to units present in published data.

In the remaining, Section 2 reviews the theory of anonymization. Also the formal notion of disclosure is constructed. Section 3 reviews DP from a view point of statistical finite population analysis. As a result, we obtain reasoning to control the level of protection brought by DP. Section 4 concludes that the current practice of official statistics can be supported by DP.

2 Modernizing the Theory of Anonymization

2.1 Statistical Disclosure Control

Confidentiality in statistics indicates that data collected by a statistical agency are to be strictly confidential and used exclusively for statistical purposes (Principle 6, Fundamental Principles of Official Statistics (A/RES/68/261 from 29 January 2014), the U.N.). The same terminology in a computer security context implies to ensure that information is accessible only by authorized parties (ISO/IEC 27000), but the present article concerns confidentiality in the statistical sense. Anderson and Seltzer (2009) describe the historical development of confidentiality notion in the U.S..

Empirical evidence such as Singer et al. (1993, 2003) supports a fact that the pledge of confidentiality promotes honest responses on a survey. Hence statistical agencies have long been developing methods for the protection of confidentiality. On the other hand, official statistics exists to reveal facts. This intrinsic nature of

statistics contradicts the protection of confidentiality. Therefore, publishing statistics confronts a tradeoff between the risk of the breach of confidentiality (disclosure risk) and the analytical validity of statistics (utility).

Endeavors to solve this dilemma constitute a field of researches called Statistical Disclosure Control or Limitation (SDC or SDL). For general understanding around SDC, readers should refer to Duncan et al.'s (2011) textbook. More recently Templ (2017) specializes in the microdata protection of SDC, and this textbook leads to hands-on understanding of important concepts by supplied codes based on R packages: `sdcMicro` and `simPop`.

A significant part of SDC consists of contributions from statistical agencies; see Doyle et al. (2001) or Hundepool et al. (2012) for example. This fact characterizes SDC in two ways. First, SDC is practical, sticking closely to the real issue of a statistical agency. Willenborg and de Waal's (1996, 2000) lecture notes represent this nature. Second, since statistical agencies are amongst the earliest adopters of a relational database, they have been promoting collaboration between statisticians and computer scientists. This collaboration produces, e.g., an enlightening volume edited by Nin and Herranz (2010). Also a series of biennial proceedings of Privacy in Statistical Databases after Domingo-Ferrer and Tora (2004) symbolizes the long-lasting cooperation.

2.2 SDC among Privacy Researches

The practical nature of SDC, however, tends to limit a view point to being field specific. It is important to recognize that privacy researches outside official statistics are related to SDC.

Conceptualizing privacy is an ongoing research issue. Readers interested in fundamental arguments on privacy should refer to Solove (2008). Terminology on privacy properties is proposed by Pfitzmann et al. (2010), and Deng et al. (2011) consider some of them including unlinkability, anonymity and pseudonymity, plausible deniability, undetectability and unobservability, and confidentiality. Their argument from the view point of secure software engineering should be intriguing to many statisticians.

Since the 1970s the law provides people with control over their data to protect privacy. According to Solove (2013), this "privacy self-management" approach is so widely accepted, although it may be no longer sound enough. Lowrance (2012) states that confidentiality serves privacy, which is in the sense of self-management. In fact, a sense of control over one's information promotes participation in a business service or a survey, similarly to the pledge of confidentiality; see Stewart and Segars (2002) for empirical evidence.

Because social regulations have been relying on privacy self-management, concerns on the control of information, which is the goal of SDC, are prevailing amongst fields. For example, health research needs to protect private data. El Emam and Arbuckle (2013) provide a technical reference in this field, which cites many results of SDC. Dennis (2000) otherwise composes a non-technical guide for medical practitioners to protect privacy. We confirm from this work that a practice never stands without institutional limitations such as laws, regulations, guidance and governance. Hence it is natural for SDC to restrict itself within statistical institutions, but there is actually a common part among institutions. From

Lowrance's (2012) international comparison on the institution of health research, although some of its portion is inevitably outdated, we learn that the identification of an individual is universally unacceptable. As in SDC, anonymization or de-identification is of primary importance.

Computer scientists too share our interest in the control of information, but they are relatively free from institutional limitations. For example, they do not necessarily deem a statistical agency as a trusted curator of data. This point of view is hard to emerge within SDC. The technique of secure multi-party computation facilitates analyzing distributed data without a curator to conduct a survey; see, e.g., Muralidhar et al. (2016). In statistics the randomized response method (Warner, 1965) presumes the curator of data untrusted. Warner (1971) himself notes that a randomized response method serves SDC; the disclosure of confidential information can be avoided by reporting only the sum of a true value and a random value from a known distribution. A randomized response model is adopted by Google's RAPPOR to prevent a privacy breach (Erlingsson et al., 2014), which is frequently mentioned by computer scientists.

Deng et al. (2011) discriminate between hard privacy and soft privacy. The former aims data minimization, based on the assumption that personal data should not be divulged to third parties. The latter aims to provide data security and process data with specific purpose and consent, based on the assumption that a data subject is unable to control personal data and has to trust a data curator.

Hard privacy is pursued in, e.g., cryptography, reducing needs to trust other entities. On the other hand, official statistics seeks soft privacy. It is very important to recognize that schemes for hard privacy such as DP lead to lower utility since it is stronger than soft privacy.

Realized privacy breaches such as AOL's scandal on publishing web search queries (Barbaro and Zeller, 2006) depict the need of privacy researches off a statistical agency. Even though some of them are impractical and unreal, they consist of fertile soil. After Agrawal and Srikant (2000) and Lindell and Pinkas (2000), the field of Privacy Preserving Data Mining (PPDM) blooms in computer science. PPDM enables data mining while controlling disclosure. A volume of reviews by Aggarwal and Yu (2008) is very informative on this field. See also Mendes and Viela (2017) for a newer review.

PPDM assures the validity of data analysis in only designed cases, by which disclosure is easier to control. An example queries a database interactively. Dwork (2011) puts "the advantage of the interactive approach is that only the questions actually asked receive responses." However, exploratory data analysis (Tukey, 1977) absolutely needs published data.

A branch of PPDM called Privacy Preserving Data Publishing (PPDP) undertakes to disseminate data in a private manner. This goal is similar to that of a statistical agency, but the most essential difference seems the scope of data types. A statistical agency traditionally publishes tables and microdata. On the other hand, PPDP deals with more types including transaction data, trajectory data, social networks, and textual data; see Fung et al. (2011) for these contexts.

Table 1 Conditions of identification (Marsh et al., 1991)

(a)	Published data are measured consistently with filed key variables (i.e., no misentry or misclassification, etc.).
(b)	Published data contain the target.
(c)	The target is a population unique. That is, no other entity in a population has the same attributes on key variables as those of the target.
(d)	The population uniqueness of the target is ascertained.

2.3 Spontaneous Reform of SDC

The principal hardship to publish nontraditional data is that it may be unrealistic to assume that attackers or adversaries, who want to identify an individual, know only a limited number of attributes of targets. These known attributes are called key variables or quasi-identifiers (Dalenius, 1986), discriminating from direct identifiers such as a name or an address.

In many countries including Japan, publishing official statistics legally requires the unidentifiability of survey respondents, which is not the same concept as confidentiality. This difference, however, enables data users to statistically estimate confidential or sensitive variables: A sensitive variable is regarded as unknown in the following, although non-sensitive variables can be unknown.

The well-established rationale of publishing official statistics employs a logic that an individual can not be identified among multiple individuals of the same attributes on key variables. It is assumed that an attacker files identified individuals whose attributes are partly known, and searches published data for those who have the same attributes as those of a filed individual. Then Marsh et al. (1991) define the success of the identification of a filed individual, or a target, as the product of 4 conditions summarized in Table 1. Actually, masking for de-identification or anonymization aims to preclude at least one of these 4 conditions. For instance, k-anonymity (Sweeney, 2002) is sufficient for population uniqueness (Condition (c)) to fail, since published data are masked so that multiple records have the same attributes.

It is worthy of note that population uniqueness (Condition (c)) must be verified (Condition (d)) for the success of identification, since Condition (d) is often neglected. The author considers that the verification of population uniqueness needs the complete frame of a subpopulation that includes the target to be verified (Hoshino, 2016). For example, to verify the uniqueness of an attorney, an attacker needs the whole list of attorneys, which is only a part of a population. However, completing the frame of a subpopulation is not always easy. We should bear the likelihood of Condition (d) in mind for assessing identification risk.

The key variables of official statistics are often regarded as basic demographic variables, which are not many. Then population uniques can be limited enough to declare that individuals are *de facto* unidentifiable. Rocher et al. (2019) point out that the likelihood of population uniqueness (Condition (c)) may be correctly evaluated even if population uniques are limited, but few population uniques imply low likelihood for ordinary people to find their acquaintances, i.e., filed people, in a published data set (Condition (b)). The likelihood of identification is the product of the likelihoods of the 4 conditions, which can be low even though a part of them is high. Disclosure control is often tailored for ordinary people in official statistics.

A data broker who has the full information of people is not necessarily assumed in order to provide useful data; see Brandt et al. (2008, p.140) for example.

The reasoning above essentially needs the limited selection of key variables. If key variables are so many that almost all records are population uniques, then ordinary people would identify their neighbor in published data. The situation was different from publishing, e.g., a long sequence of web search queries, which can be many key variables to identify an individual, and the assumption of limited key variables has been mostly plausible concerning official statistics.

However, social scientists urge to redefine the practice of official statistics, as once their demand opened their access to microdata. They want more richly detailed data of individuals, firms and other organizations. In particular, longitudinal or panel surveys, which contain a time series of key variables, are unprecedentedly required by researchers. The cost of longitudinal surveys can delay the reform of official statistics, but linking or matching records of the same entity among different files can satisfy researchers' demand at relatively little cost.

Linked data can be produced even if the same entity is not included among different files. Statistical matching, distinguished from exact matching, estimates the unobserved attributes of an entity, and the estimates are linked to its observed attributes. Those interested in statistical matching should refer to D'Orazio et al. (2006).

At the frontier of social sciences, researchers increase access to linked data from multiple surveys and administrative records through cumbersome administrative and legal arrangements (Butz and Torrey, 2006). Moreover, the use of linked data often requires traveling to a secure on-site data center, which is called "on site" in Japan (Nakamura, 2017), or submitting statistical software to a remote analysis system. O'Keefe (2015) mentions some examples of restricted accesses of official statistics.

The reason of these inconveniences for researchers is apparently the high risk of identification inherent in linked data. Firstly, linked data can be valuable in business. For example, Facebook links its own data with that of a data broker (Datalogix) to market fish oil (Goel, 2014). The rise of economic value attracts more attackers. Secondly, the number of key variables monotonically increases by linking records, which implies many population uniques. Rocher et al. (2019) claim that 99.98% of Americans would be correctly identified using 15 key variables. Sweeney (2000) guesses that 87% of Americans are likely to be unique with respect to only 3 key variables: {5-digit ZIP, gender, date of birth}. Geographic information such as a ZIP code is widely regarded as a strong key variable, and thus it is swapped in producing Anonymized Data of the Japanese census for example. However, we should note that Sweeney's guess employs the pigeonhole principle in an unconvincing way.

Consequently, as National Research Council (2007) put, linked data are usually not publically available due to confidentiality concerns. Using linked data is generally allowed within a "safe" situation, which is ensured not only by statistical controls but also by managerial controls. Ritchie (2008, 2017) conceptualizes provision for data access as Five Safes of control dimensions: safe projects, safe people, safe data, safe settings and safe outputs.

Nevertheless, restricting data access impedes researches. It also consumes non-negligible resources of statistical agencies, which implies that they may have disincentive to promoting researches (Fienberg, 2005). Because safe public-use data

can solve these issues, efforts to produce such linked data have been observed. The most prominent one seems the Survey of Income and Program Participation (SIPP) Synthetic Beta, which is a U.S. Census Bureau product that links records from a household survey with administrative tax and benefit data; see Benedetto et al. (2018) for further details. This ongoing project employs the approach of synthetic data initiated by Rubin (1993).

2.4 Synthesis for Publishing Unsafe Data

The original idea of synthetic data derives from the solid theory of multiple imputation for missing values; see a classic textbook by Rubin (1987). Adapted for SDC, it regards the unobserved part of a population as missing. Then missing values are randomly imputed several times in order to express the uncertainty of imputation, and from each imputed population random samples are drawn to be published. Drechsler (2011) provides an overview of this context.

Synthetic data can consist of records that have no correspondence to real entities. Then the identification of a record has no direct sense, and this type of data should be less difficult to publish. However, publishing only random values, which is called fully synthetic approach, may overprotect data. By limiting synthesis only to an unsafe part, the utility of published data can improve. This approach is known as partial synthesis (Little, 1993). It is beneficial also because of the less burden of modeling a population. Statistical products such as the SIPP Synthetic Beta employ partial synthesis. As regards the estimation of the uncertainty of synthetic data analysis, Raab et al. (2017) state that distinction between fully and partially synthetic data is not meaningful, though.

Aforementioned statistical matching is closely related to imputation. It also estimates unobserved attributes of entities, but it does so generally with less information on association between observed and unobserved variables. Imputation in a single file can exploit observed association among variables of the same entity, which is usually impossible for statistical matching since a different file contains different entities. Statistical matching is yet another view of synthetic data.

Beckman et al. (1996) propose the use of a synthetic population or a pseudo-population in microsimulation or agent-based modeling. Heard et al. (2015) review these literatures in economics, finance, ecology, biology and epidemiology, where researchers simulate actions and interactions of entities within a system or a population to gain insight into a complex system. Regarding SDC, Quatember (2015) mentions that the effect of data masking can be evaluated by sampling from a pseudo-population.

Templ et al. (2017) review main approaches to the generation of a pseudo-population. According to them, the most frequently used method is sampling from an empirical distribution (i.e., observed proportions) with the restriction of marginal population frequencies given by census tables. Joint population frequencies are commonly estimated by Iterative Proportional Fitting (IPF, Deming and Stephan, 1940) or raking; see Bishop et al. (1975) for the understanding of IPF.

Unfortunately, census tables are not always available. Then population frequencies need to be estimated. Quatember (2015) relies on the Horvitz-Thompson

(HT) estimator, which multiplies a sample frequency by the inverse of its inclusion probability; see Horvitz and Thompson (1952). In other words, an empirical distribution is adjusted reflecting survey weights.

However, when an inclusion probability is low, the HT estimator is useless notably for small population frequencies because of its large variance. This issue matters to linked data because they generally consist of small frequencies: Samples are scattered in a relatively high dimensional space. Many statistical applications suffer from this phenomenon, which is called Large Number of Rare Events (LNRE) by Khmaladze (1987). Baayen (2001, Section 2.4) provides the numerical examples of LNRE in linguistic. Even the frequency of a population frequency (Good, 1953) is not a modest objective to estimate. Its unique unbiased estimator is useless as Shlosser (1981) explains. In other words, sampling variations conceal population uniqueness.

This is the reason why population uniques are estimated by various models since Bethlehem et al. (1980). A model supplies auxiliary information to stabilize the estimation with the loss of unbiasedness. For example, Rocher et al. (2019) claim that the likelihood of population uniqueness for each record can be better estimated using correlation among attributes. Their estimated likelihood, however, never reflects difference between a model and a true population. This deterministic difference obstructs any model to rigorously prove population uniqueness.

The large stochastic difference between samples and a population can be exploited to improve utility and safety simultaneously. Fienberg (1994) proposes to publish bootstrap (Efron, 1979) samples from an empirical distribution smoothed by auxiliary information such as census tables, past surveys and nonsampling errors (from editing, matching, nonresponse, etc.). This smoothing works as masking that affects Condition (a) of Table 1, and published data provide better inference on a population. Moreover, replicates by bootstrapping enable us to measure a between-replicates variation as multiple imputation; see Fienberg et al. (1998) for a more detailed argument on this approach.

The smoothing of an empirical distribution is a type of population modeling. It should be of better utility if it more resembles a population in which the users of data are interested. This principle is widely accepted for generating synthetic data; Snoke et al. (2018) measure the utility of synthetic data by distributional similarity between the population of raw data and a model used to generate synthetic data. Although they state that the risk measure of synthetic data is under development, it is reasonable to measure that risk using the utility measure reversely. As stated at the very beginning, disclosure risk mirrors utility.

Actually, similarity between raw data and published data is frequently used to evaluate the disclosure risk of perturbation or untruthful masking such as additive noise (e.g., Brand, 2002). Most of perturbation is random, where average similarity is usually evaluated, and this averaging measures distributional similarity between raw data and a model used to generate perturbed data.

The above utility measure of synthetic data examines the population of raw data instead of raw data, which results from the difference of the population of interest. A utility measure concerns a data user. Hence the examined population has to be one in which a data user is interested for analysis. By contrast, a risk measure concerns an attacker who tries to identify a record. Then the examined population can be raw data, which are the population for attackers who know (a part of) survey respondents. If survey respondents are assumed to be unknown,

the population should be the sampling frame of raw data. This is the case of traditional risk measures based on population uniques.

We should note that the estimation of the sampling frame of raw data is equivalent to the estimation of population frequencies, which is difficult as stated before. Especially in the case of linked data, the curse of dimensionality arises. That is, raw data are too sparse to estimate a high dimensional population as Gottschalk (2004) points out. Difficult estimation leads to the uncertain measurement of risk, and thus it is understandable to measure an object that does not require estimation. This is another reason why a sampling frame is usually neglected in the risk measurement of perturbation.

To summarize, the risk of synthetic data should be measured by the distributional similarity between the population of raw data and a model used to generate synthetic data, where the population is set depending on whether an attacker knows survey respondents or not.

This way integrates random perturbation and random synthesis. The following Example 1 demonstrates that common discrimination between synthesis and perturbation is meaningless. Fully synthetic data can be safe not because they do not contain real entities but because they can be very different from the population of raw data. All kinds of random modification of raw data such as synthesis, perturbation and subsampling require a unified treatment as random masking.

Example 1 *Let us publish the height of an individual: 152 cm. Synthesis may replace this value with a sample from the normal distribution with mean 152 and variance 1, which is an infinite population obtained by smoothing the datum. Perturbation may add a noise subject to the standard normal distribution to the datum. The both cases of results have the same distribution.*

2.5 Formal Notion of Disclosure

We now reconsider identification. Identification is harmful when it entails identity disclosure, which is explained as "(it) occurs when a data subject is identified from released data" by Duncan et al. (2011, Glossary). What occurs is that the sensitive variables of a unique entity take one combination of values with probability one. In other words, identification restricts the conditional distribution of sensitive variables on a point, given key variables. This degeneration of the conditional distribution occurs since only one entity is distributed. Therefore, we employ the following Definition 1 of identity disclosure.

Definition 1 *The identity disclosure of a target occurs if and only if the following two properties of the population of the target are shown: (I) the conditional distribution of sensitive variables given that key variables take the values of the target and (II) its size is one.*

In Definition 1, "size" expresses the number of entities that have the same attributes as those known of the target; see the following Example 2.

Example 2 *Let us consider publishing a data set of 3 variables: Sex, Height and Age. For simplicity, they all are dichotomous as {F, M}, {T, S} and {O, Y}, respectively. Then the sample space or the set of all possible combinations of attributes is*

$$\{(F,T,O),(F,T,Y),(F,S,O),(F,S,Y),(M,T,O),(M,T,Y),(M,S,O),(M,S,Y)\} = \Omega.$$

The element of Ω is called a point. Population frequencies on these points are denoted by $n_{FTO}, n_{FTY}, n_{FSO}, \ldots, n_{MSY}$, which sum up to n.

For an attacker who does not know the target's sex, height and age, identity disclosure is equivalent to know (i) the unconditional distribution of the population, which is relative frequencies over Ω: $(n_{FTO}/n, n_{FTY}/n, n_{FSO}/n, \ldots, n_{MSY}/n)$, and (ii) $n = 1$.

In this case, (ii) implies that only one point of Ω can have a positive probability. Hence the distribution of the attributes degenerates, and thus knowing (i) maps the target to the attributes of, say, (F, T, O). Also (ii) implies that the attributes of (F, T, O) is inversely mapped to the target. This bijection between the population of the target and $\{(F, T, O)\}$ is the exact sense of identification under Definition 1.

Next we consider an attacker who knows the target's Age=O. Then identity disclosure is equivalent to know (i') the conditional distribution given Age=O or $(n_{FTO}/(n_{FTO} + n_{FSO} + n_{MTO} + n_{MSO}), n_{FSO}/(n_{FTO} + n_{FSO} + n_{MTO} + n_{MSO}), n_{MTO}/(n_{FTO} + n_{FSO} + n_{MTO} + n_{MSO}), n_{MSO}/(n_{FTO} + n_{FSO} + n_{MTO} + n_{MSO}))$, and (ii') $n_{FTO} + n_{FSO} + n_{MTO} + n_{MSO} = 1$.

Similarly, (ii') implies that the conditional distribution given Age=O degenerates. Also implied is the bijection between the population of the target and the subspace of Ω.

When an attacker's target is not included in published data, Condition (b) of Table 1 is not satisfied and thus the identification of the target does not happen. Hence fully synthetic data are claimed to be safe. However, even if the target is not included in published data, the conditional distribution of sensitive variables and its size may be correctly estimated. It is noteworthy that the population by definition contains the target.

Irrespective of the presence of the target in published data, identity disclosure in our sense can occur. The following Example 3 demonstrates this possibility, where the estimation of the size is not statistical, though.

Example 3 *Let us synthesize the raw data of {152 cm, the host and co-executive producer of Fresh Air}. Suppose that 10000 synthetic records of the host and co-executive producer of Fresh Air are published, in which each height is a sample from the normal distribution with mean 152 and variance 1. Then observing these synthetic data, an attacker who knows only the occupation of Terry Gross would guess her height rather correctly at the average of 10000 heights. This is because the number of individuals who are the host and co-executive producer of Fresh Air is one in a population of human beings and the conditional distribution of height given that occupation is almost revealed.*

Dwork et al. (2017) review a tracing attack, where an attacker wants to determine if the target is present in published data or not. As they claim, mere presence in data can be highly sensitive information, but presence itself is not disclosive. It causes disclosure only when inference on unknown information follows. Terry Gross loses nothing even if her presence in the nominal list of living females is correctly determined.

Closely related to identity disclosure, attribute disclosure is explained as "the disclosure of information about a population unit without (necessarily) the identification of that population unit *within a data set*" by Duncan et al. (2011, Glossary). It implies that the conditional distribution of sensitive variables degenerates, where the size is not necessarily one.

In Example 2, the unconditional distribution of the attributes can degenerate even if $n > 1$. Then the attacker is certain about the attributes of the target because any member of the population of the target is mapped to a unique combination of attributes (surjection). Its inverse mapping, however, is defined if and only if $n = 1$. In other words, an entity who has the unique attribute may not be the target when $n > 1$. We note that the size condition is necessary for identification.

Therefore, we employ the following Definition 2 of attribute disclosure, which includes Definition 1 as a special case. It is noteworthy that attribute disclosure can also occur without the presence of the target in published data.

Definition 2 *The attribute disclosure of a target occurs if and only if the following two properties of the population of the target are shown: (I) the conditional distribution of sensitive variables, given that key variables take the values of the target and (II) it degenerates (i.e., takes one combination of values with probability one).*

Clinging to the notion of presence impairs our understanding of disclosure. Actually presence is irrelevant to disclosure. Therefore, presence control is indirect and inefficient in disclosure control. A direct objective to control disclosure is the certainty of population frequencies.

Let us consider Example 2 again. For an attacker who knows the target's Age=O, identity disclosure is uncertain if it is uncertain whether $n_{FTO} + n_{FSO} + n_{MTO} + n_{MSO} = 1$. Similarly, for an attacker who knows the target's Age=O and Height=T, identity disclosure is uncertain if whether $n_{FTO} + n_{MTO} = 1$ is uncertain. Generalizing this consideration, if every population frequency over Ω is uncertain about uniqueness, then identity disclosure is uncertain for any attacker.

Analogously, certainty about a population frequency being zero causes attribute disclosure because the population frequency of complemental attributes is necessarily positive, thus degenerates, owing to the fact that the population is not empty. Therefore, every population frequency must be uncertain about zero for no attribute disclosure to be certified. .

Certainty about a large population frequency may not seem problematic, but differencing from outer information of a population frequency, which may be provided by census tables, could result in disclosure. This issue is fundamental to the protection of contingency tables; see a methodological introduction by Giessing (2004) to the famous software of tabular data protection: τ-ARGUS. Contingency tables provide marginal frequencies, which are simultaneous equations that solve the range of the frequency of a cell. Therefore, when many marginal frequencies are available, the ranges of frequencies could be narrow enough for disclosure to occur. This sophisticated version of differencing is called reconstruction attack (Dinur and Nissim, 2003).

An important observation from tabular data literatures is that outer information may invalidate deterministic uncertainty such as an interval expression. Deterministic information enables an attacker to deduce a deterministic assertion. Because disclosure is a deterministic concept, introducing stochastic uncertainty to forestall deduction is a good methodology in principle. We do not have to worry about infinite possibilities of types of outer information, provided that the uncertainty of a population frequency is furnished by random masking.

In conclusion, every population frequency, regardless of its value, ought to be stochastically uncertain to prevent definite disclosure in any occasion. This objective does not depend on the selection of key and sensitive variables.

The notions of key variables and sensitive variables are very much criticized as subjective, vulnerable to misselection and so on. According to Fung et al. (2010), the selection of key variables is regarded as an "open problem."

However, assuming key variables is advantageous to publish useful data. If a specific selection of key variables is valid, then it suffices to make a part of population frequencies uncertain. We should not waste knowledge on attackers or outer information. In order to select key variables objectively, Hoshino (2016) proposes to estimate a high percentile of the attackers' distribution over the ability of identification, which is expressed as the combination of key variables that he or she can use.

Now we set our notation to formally state a unified treatment of random masking. It is worthy of note that deterministic masking such as recoding (generalization) or suppression collapses the sample space.

Suppose that the sample space consists of J points, each of which is indexed by j. The population frequency of the jth point is n_j, and $\mathbf{n} := (n_1, n_2, \ldots, n_J)$. The sum of population frequencies is $n := \sum_{j=1}^{J} n_j$. The frequency of the jth point in published data is m_j, and $\mathbf{m} := (m_1, m_2, \ldots, m_J)$. The sum of published frequencies is $m := \sum_{j=1}^{J} m_j$.

Our definition of disclosure is valid for a continuous sample space. However, we only consider the discrete case, which covers real data. Also, taking $J \to \infty$ approximates the continuous case; see an example of this limiting argument by Hoshino (2009).

A microdata set is equivalent to the set of frequencies over the sample space, and we are considering to publish not a frequency table but microdata. This issue is closely related to histogram publishing known in PPDP; see a statistical paper by Wasserman and Zhou (2010). Histogram publishing allows m_j to be a real number, but microdata publishing needs m_j to be a nonnegative integer. Differentially private histogram publishing by Ghosh et al. (2012) applies an additive noise from the discrete Laplace distribution (Inusah and Kozubowski, 2006) to n_j. Although m_j can then be a negative integer, yet censoring at zero remedies this issue. Statisticians should refer to Rinott et al. (2018) for more information.

A random masking is the distribution of $\mathbf{m}$ whose support is

$$\mathcal{M} := \{\mathbf{m} : m_j \in \{0, 1, \ldots, m\}, j \in \{1, 2, \ldots, J\}, \sum_{j=1}^{J} m_j = m\}.$$

The parameter space of $\mathbf{n}$ is denoted by

$$\mathcal{N} := \{\mathbf{n} : n_j \in \{0, 1, \ldots, n\}, j \in \{1, 2, \ldots, J\}, \sum_{j=1}^{J} n_j = n\}.$$

Regarding $\mathbf{n}$ as a parameter vector stems from the traditional view of finite population sampling; see, e.g., Godambe (1955).

DP is a general notion to evaluate random masking, and the next section considers DP in the framework of statistical estimation. Actually, DP bounds the variance of the estimator of a parameter. As we can see in the following Example 4, the statistical theory tells us the accuracy of the estimation of n_j.

Example 4 *Let us assume that an attacker knows survey respondents. Then* $\mathbf{n}$ *becomes the frequency vector of raw data. Its bootstrap samples* $\mathbf{m}$ *are subject to the multinomial distribution* $\mathrm{P}(\mathbf{m};\mathbf{n}) = m!\prod_{j=1}^{J}(n_j/n)^{m_j}/m_j!$. *In this case,* m_j *is sufficient for* n_j*, and thus we need to consider only the marginal distribution of* m_j *to evaluate the correctness of the estimation of* n_j*. The margin of* m_j *is subject to the binomial distribution* $\mathrm{P}(m_j;n_j) = m!(n_j/n)^{m_j}/m_j!(1-n_j/n)^{m-m_j}/(m-m_j)!$. *If* $m=n$*, then* m_j *is the unbiased estimator of* n_j*. It follows* $\mathrm{V}(m_j)/n_j = (1-n_j/n)$*, which increases as* n_j *decreases. In this sense, bootstrapping is more protective for smaller* n_j*.*

3 Differential Privacy

3.1 Masking for Apprentices

Numerous privacy notions have been proposed by computer scientists, neglecting the utility of data. Some renowned examples are given below. k-anonymity (Sweeney, 2002) damages the data of isolated entities seriously. ℓ-diversity (Machanavajjhala et al., 2007) prohibits the conditional distribution of sensitive variables given key variables to degenerate. It prevents attribute disclosure in the sense of Definition 2, but the subsequent statistical analysis of sensitive variables should be heavily biased. t-closeness (Li et al., 2007) masks data so that key variables and sensitive variables are almost independent. This mask completely destroys statistical analysis because researchers use data in order to know relationship between demographic (key) variables and unknown (sensitive) variables.

Variants of these notions mask data until some syntactic condition is met, so that the ability of an attacker to link key variables to sensitive variables is restricted. Therefore, Clifton and Tassa (2013) call these notions syntactic models.

Various attacks have been proposed to defeat those syntactic models, and a loser develops another model, which invites another attack. The repeat of this updating strengthens the definition of privacy. According to Dwork et al. (2017), DP "was first proposed in 2006 and so far has not required strengthening." What is so far required is weakening.

Syntactic models protect given data, which is orthodox in SDC. On the other hand, DP is the property of a random masking. This way is more akin to that of a frequentist who is interested in the long-run property of a random variable.

Let us see Definition 3 of DP below, where $\mathcal{A}$ does not depend on a specific D since this dependence can be decisive to infer D from realized $A(D)$.

Definition 3 (Dwork, 2006) *Let* D *be a data set. The space of* D *is denoted by* $\mathcal{D}$*. The random masking of* D *is denoted by* $A(\cdot)$*. The support of* $A(\cdot)$ *is denoted by* $\mathcal{A}$*.* Δ *denotes a unit change. A random masking* $A(\cdot)$ *is* ϵ*-DP (Differentially Private) if and only if*

$$\mathrm{P}(A(D+\Delta)\in S)/\mathrm{P}(A(D)\in S) \le \exp(\epsilon) \tag{1}$$

for all $S\subset\mathcal{A}$*, for all* $D\in\mathcal{D}$ *and for all* Δ *such that* $D+\Delta\in\mathcal{D}$*.*

The left hand side of eq. (1) is regarded as the evidence of a unit change of D. As $\mathrm{P}(A(D+\Delta)\in S)$ departs from $\mathrm{P}(A(D)\in S)$, the unit change Δ should be easier to detect by observing masked results. Hence the left hand side of eq. (1) is bounded around one to conceal Δ. It might be noteworthy that no random masking

is ϵ-DP when ϵ is negative, since by definition $D + \Delta$ and D must be exchangeable in eq. (1). Thus in literatures $\epsilon > 0$. As ϵ increases, Δ is more revealed, and the tuning parameter ϵ in eq. (1) is called a privacy budget.

Observing Definition 3, we note that any random masking with the same likelihood function up to a constant multiplication provides the same evidence of a unit change. This irrelevance of statistical evidence to any structure that does not affect likelihood is called the likelihood principle; see Birnbaum (1962). Accordingly, DP is irrelevant to whether random masking is synthesis, perturbation or subsampling. This is the unified treatment of random masking motivated by Example 1.

Eq. (1) is required to hold for the most detectable case of $D \in \mathcal{D}$. This minimization of the maximum information enables even an apprentice to mask data. It contrasts with the current practice of SDC, which needs an artisan who tailors masking depending on D. Expelling artisanship, however, costs utility very much. Therefore, the essential idea of Soria-Comas et al. (2017) is to set $\mathcal{D} = \{D\}$. Likewise, many variants of DP have been proposed to escape from the minimax reasoning of the original DP.

Approximate DP (Dwork et al., 2006a) requires for positive δ

$$\mathrm{P}(A(D + \Delta) \in S) \leq \exp(\epsilon)\mathrm{P}(A(D) \in S) + \delta \tag{2}$$

instead of eq. (1). Obviously, approximate DP is equivalent to DP when $\delta = 0$. Approximate DP allows exceptions to eq. (1) for small $\mathrm{P}(A(D + \Delta) \in S)$. This dependence on the absolute value of $\mathrm{P}(A(D + \Delta) \in S)$ implies that δ has to be tuned depending on the width of the parameter space, which is quite artisanal. Exceptions to eq. (1) can be controlled stochastically in many senses as Meiser (2018) discusses. Otherwise Nissim et al. (2007) consider ϵ-DP noise that depends on D, where the amount of noise is carefully designed not to reveal D, while Dwork (2011) states "To satisfy differential privacy, the noise must be independent, not only of the true answer, but also the size of the database."

More relaxations of DP develop by considering the interpretation of DP, since they have to protect privacy in some sense. Typical interpretations are reviewed in the next section.

3.2 Semantics of DP

What is protected by DP? The answer to this question depends on the interpretation of the unit change Δ in eq. (1). Dwork (2006) interprets Δ as the inclusion (or exclusion) of a target. According to Dwork (2011), a new privacy goal is to " minimize the increased risk to an individual incurred by joining (or leaving) the database."

This original interpretation derives from the cryptographic interest in presence, which is irrelevant to disclosure as we have seen. Also Δ can be the inclusion of someone with the same attributes as those of the target. Then the left hand side of eq. (1) is not even the evidence of the presence of the target in published data. Nevertheless, the original interpretation dominates in the literature. The notion of presence again impairs our understanding.

When we regard D as a parameter (vector) and $\mathcal{D}$ as its parameter space, the left hand side of eq. (1) is the Likelihood Ratio (LR) of Δ to its nonexistence. This form suggests the hypothesis test of a unit change.

The composite hypothesis of the existence of some Δ can be tested simply between the most indistinguishable hypotheses of the existence and the nonexistence. Also the Neyman-Pearson lemma (Neyman and Pearson, 1933) assures that the most powerful test rejects the null hypothesis if the LR is larger than some threshold. Therefore, in eq. (1) the LR is bounded by $\exp(\epsilon)$ so that the null hypothesis of no unit change is never rejected.

This view is a finite population version of the popular interpretation of DP that assumes an infinite population from which D is sampled. For example, Wasserman and Zhou (2010) regard D as i.i.d. samples. In the context of DP, hypothesis testing under an infinite population is well studied by Liu et al. (2019); see also references therein. Often D is assumed to have a distribution $f(\cdot;\theta)$, where θ is a parameter (vector) and $\mathcal{D}$ is a *sample* space. Bayesians need a prior for θ, and thus $\mathcal{D}$ can not be a parameter space for them.

In the Bayesian context, the LR is the Bayes factor. It is used by Jeffreys (1935, 1961) for the purpose of hypothesis testing to evaluate the evidence of a scientific theory. This Bayes factor view produces other variants of DP. For example, Kifer and Machanavajjhala (2014) regard a prior for θ as an attacker's background knowledge. This is another way to protect D at hand in $\mathcal{D}$; they justify the assumption of an informative prior due to the no free lunch theorem (Kifer and Machanavajjhala, 2011) that one can not ensure privacy and utility simultaneously without making assumptions about an attacker's background knowledge. Also in Kifer and Machanavajjhala (2014), $D + \Delta$ and D are mutually exclusive 2 secrets, and Δ is not a unit change or adjacency anymore.

A different interpretation of DP arises by rewriting eq. (1) as

$$\frac{\log \mathrm{P}(A(D+\Delta) \in S) - \log \mathrm{P}(A(D) \in S)}{|\Delta|} \leq \epsilon, \tag{3}$$

where $|\Delta| = 1$. The left hand side of eq. (3) corresponds to the gradient of the log likelihood function, where D is a parameter (vector). For a continuous parameter space $\mathcal{D}$, we can take the limit of $|\Delta| \to 0$. Then under regularity conditions, eq. (3) bounds the (partial) derivative of $\log \mathrm{P}(A(D) \in S)$ in the limit. This idea leads to the notion of derivative privacy (Hoshino, 2018). Derivative privacy naturally limits the Fisher information of D; the next section demonstrates the discrete version of this argument.

Smith (2008) considers the maximum likelihood estimator of θ under DP. His asymptotic argument also depends on the Fisher information.

3.3 Bounding the Accuracy of Population Frequency Estimation

The gradient of a likelihood function decides the accuracy of parameter estimation. Hence when it is bounded by DP, the variance of the estimator of a parameter is inevitably bounded. This section explicates such a bound of the unbiased estimator of a population frequency.

We use notation introduced in Section 2.5 of a finite population. That view is advantageous to discern identification since the resolution of $\mathcal{D}$ decides the possibility of the identification of an entity. Accordingly, Δ has to be the move of an entity in a population as in Dwork et al. (2006b).

First we state the DP condition (1) in our case. A random vector $\mathbf{m}$ is ϵ-DP, if and only if for all $\mathbf{m} \in \mathcal{M}$, for all $\mathbf{n} \in \mathcal{N}$ and for all $\mathbf{n} + \Delta \in \mathcal{N}$,

$$\frac{\mathrm{P}(\mathbf{m}; \mathbf{n} + \Delta)}{\mathrm{P}(\mathbf{m}; \mathbf{n})} \leq \exp(\epsilon). \tag{4}$$

Second, we express the Hammersley-Chapman-Robbins inequality as Theorem 1 below; see Lehmann and Casella (1998, p.113). There $\mathcal{M}$ is assumed to be common between the cases of $\mathbf{n}$ and $\mathbf{n} + \Delta$ as Definition 3 of ϵ-DP requires.

Theorem 1 (Hammersley, 1950, Chapman and Robbins, 1951) *Suppose that* $\mathbf{m}$ *is distributed subject to* $\mathrm{P}(\mathbf{m}; \mathbf{n})$*, and* $\mathrm{P}(\mathbf{m}; \mathbf{n}) > 0$ *for all* $\mathbf{m} \in \mathcal{M}$*. Let* $g(\mathbf{n})$ *be an estimand. If* $\mathbf{n}$ *and* $\mathbf{n} + \Delta$ *are two values for which* $g(\mathbf{n}) \neq g(\mathbf{n} + \Delta)$*, then for any unbiased estimator* γ *of* $g(\mathbf{n})$*,*

$$\mathrm{V}(\gamma) \geq [g(\mathbf{n} + \Delta) - g(\mathbf{n})]^2 / \mathrm{E}[\frac{\mathrm{P}(\mathbf{m}; \mathbf{n} + \Delta)}{\mathrm{P}(\mathbf{m}; \mathbf{n})} - 1]^2.$$

We note that the DP condition (4) immediately implies for all $\mathbf{n}$ that

$$\mathrm{E}[\frac{\mathrm{P}(\mathbf{m}; \mathbf{n} + \Delta)}{\mathrm{P}(\mathbf{m}; \mathbf{n})} - 1]^2 \leq (\exp(\epsilon) - 1)^2.$$

Third, we set $g(\mathbf{n}) = n_j$ and write $\gamma = \hat{n_j}$. When Δ induces the increment of n_j, $[g(\mathbf{n} + \Delta) - g(\mathbf{n})] = 1$. Then as a special case of Theorem 1, we have the following Theorem 2. Table 2 provides the lower bound of eq. (5) for some ϵ.

Theorem 2 *Suppose that* $\mathbf{m}$ *is* ϵ*-DP in the sense of eq. (4). Then for any unbiased estimator* $\hat{n_j}$ *of* n_j*,*

$$\mathrm{V}(\hat{n_j}) \geq \frac{1}{(\exp(\epsilon) - 1)^2}. \tag{5}$$

It is noteworthy that $\mathrm{P}(\mathbf{m}; \mathbf{n}) > 0$ for all $\mathbf{m} \in \mathcal{M}$ when DP holds. Considering its contrapositive, if $\mathrm{P}(\mathbf{m}; \mathbf{n}) = 0$ for some $\mathbf{m}$ then DP does not hold. This fact dismisses an important family of random masking from DP. If $\mathrm{P}(\mathbf{m}; \mathbf{n}) = 0$ for $m_j > n_j, j = 1, 2, \ldots, J$, then this random masking is called sampling without replacement. Hence we note Remark 1 below. In particular, simple random sampling without replacement is not ϵ-DP, as Shlomo and Skinner (2012) point out. They tacitly exclude sampling with replacement and claim that "a sampling scheme in which a population unique is sampled with a positive probability is not (ϵ-)DP," but this claim is inappropriate since it presumes that $\mathcal{N} = \{\mathbf{n}\}$.

Remark 1 *No sampling without replacement is* ϵ*-DP.*

We have seen that ϵ controls the lower bound of the variance of the estimator of a parameter. Therefore, ϵ should be set to an acceptable level of the accuracy of the estimation of a parameter. This direct logic seems new and simplifies complication described in the next section.

Table 2 Lower bound of $\mathrm{V}(\hat{n_j})$

ϵ	.01	.1	.5	1	2	3
$1/(\exp(\epsilon)-1)^2$	9900.4	90.4	2.38	.339	.024	.003

3.4 Complication of a Privacy Budget

Liu et al. (2019) state that "a key challenge is how to set an appropriate value" of ϵ for DP in practice. Also Reiter (2019) puts "finding understandable conceptualizations of privacy budgets and accuracy loss in practice remains an open challenge."

Some values of ϵ are actually used without firm reasoning. Examples are provided below. According to Reiter (2019), OnTheMap of the U.S. census bureau uses $\epsilon \approx 9$. RAPPOR of Google uses $\epsilon = \log(3)$; see Erlingsson et al. (2014). Tang et al. (2017) claim that Apple uses $\epsilon = 1$ or $\epsilon = 2$. Dwork et al. (2017) state that ϵ "should be thought of as a small constant no larger than 1."

There exist a few studies on the concrete selection of ϵ. Lee and Clifton (2011) assume a prior distribution over $\mathcal{D}$. Then the posterior distribution of D after observing published data is computed, and the maximum posterior probability with respect to D is bounded as a function of ϵ, being specific to the Laplace noise. The bound of the posterior probability is controlled by a threshold, where ϵ can be expressed as the function of this threshold. This method results in the casewise decision of ϵ. Essentially they interpret the absolute value of a posterior probability as a usual probability, which is hard to justify. Likelihood is the relative evidence of parameters; see Pawitan (2001, Chap. 2) for example. Liu et al. (2019) also assume a prior distribution over $\mathcal{D}$. Then they consider an attacker's capability to statistically test mutually exclusive secrets. There, precision denotes out of those predicted positive how many of them are actually positive, and recall denotes the fraction of the true positives that are labeled as positive. They quantify hypothesis testing based on the LR by Precision Recall (PR)-relation. Since DP bounds the LR by $\exp(\epsilon)$, this PR-relation can be a function of ϵ. Because PR-relation is more interpretable than ϵ itself, the selection should be easier. However, the selection of ϵ must be tailored to each application still.

The Bayes factor itself is the final evidence of hypothesis testing. Hence its grading by the originator (Jeffreys, 1961, Appendix B) is widely accepted. Table 3 reproduces its simplified version by Kass and Raftery (1995). In forensics, Evett et al. (2000) also grade the Bayes factor as the evidence of a hypothesis. Table 4 summarizes theirs for comparison.

Simply regarding $\mathcal{D}$ as a parameter space without a prior distribution over it, a frequentist should employ the standard LR test based on the asymptotic distribution derived by Wilks (1938). When Δ is one dimensional (i.e., not a move), $2\log(\mathrm{P}(A(D+\Delta) \in S)/\mathrm{P}(A(D) \in S))$ approximately has the χ^2 distribution with one degree of freedom. By the symmetric definition of DP, we can consider that Δ is significantly revealed for a large value of the LR. Equating 2ϵ to the $(1-\alpha)$ quantile of the χ^2 distribution, we have the critical value of ϵ that conceals Δ. In the one dimensional case, when α=10%, $\epsilon = 1.353$; when α=5%, $\epsilon = 1.921$; when α=1%, $\epsilon = 3.317$. These values seem very consistent with Table 3.

Table 3 Interpretation of the Bayes factor (Kass and Raftery, 1995)

$\log_{10}\exp(\epsilon)$	$\exp(\epsilon)$	ϵ	Evidence against null
0 to 1/2	1 to 3.2	0 to 1.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	1.2 to 2.3	Substantial
1 to 2	10 to 100	2.3 to 4.6	Strong
> 2	> 100	> 4.6	Decisive

Table 4 Interpretation of the Bayes factor (Evett et al., 2000)

$\exp(\epsilon)$	ϵ	Evidence against null
1 to 10	0 to 2.3	Limited evidence to support
10 to 100	2.3 to 4.6	Moderate evidence to support
100 to 1000	4.6 to 6.9	Moderately strong evidence to support
1000 to 10000	6.9 to 9.2	Strong evidence to support
> 10000	> 9.2	Very strong evidence to support

The decision of ϵ based on the χ^2 distribution has a solid basis and needs no casewise adjustment. However, it depends on a large sample theory where many i.i.d. samples of random masking are to be published, which is usually not the case. Also it is unclear whether the detection of Δ results in disclosure.

3.5 Privacy Budget as a Function of the Deniability of Disclosure

In this section we derive our method to select ϵ, based on Theorem 2. Ours needs no tailoring of ϵ to each application without asymptotics. Also its implication to disclosure is explicit.

We have seen in Section 2.5 that if $\hat{n_j}$ is uncertain for all j, then both identity disclosure and attribute disclosure are uncertain. We measure this uncertainty by the probability of the incorrect estimation of a population frequency.

An estimate of $\hat{n_j}$ can be a real number, but we know that true n_j is a non-negative integer. Hence an attacker should guess that n_j is the nearest integer of $\hat{n_j}$. If $|\hat{n_j} - n_j| \geq 1/2$ then the nearest integer of $\hat{n_j}$ is not the truth, and thus $\mathrm{P}(|\hat{n_j} - n_j| \geq 1/2)$ is the probability of the incorrect estimation of n_j.

Consequently, as the upper bound of this probability decreases, it is more likely that an attacker's estimation is correct. This upper bound is given by Chebyshev's inequality as $\mathrm{P}(|\hat{n_j} - n_j| \geq 1/2) \leq 4\mathrm{V}(\hat{n_j})$, where $\hat{n_j}$ is an unbiased estimator of n_j. If $\hat{n_j}$ has the least variance given by eq. (5),

$$\mathrm{P}(|\hat{n_j} - n_j| \geq 1/2) \leq \frac{4}{(\exp(\epsilon) - 1)^2} =: \alpha_\epsilon. \tag{6}$$

When $\alpha_\epsilon \leq 1$, the inequality of (6) is sharp because of the following case:

$$\hat{n_j} = \begin{cases} n_j + 1/2 & \text{with probability } \alpha_\epsilon/2, \\ n_j & \text{with probability } 1 - \alpha_\epsilon, \\ n_j - 1/2 & \text{with probability } \alpha_\epsilon/2. \end{cases}$$

Therefore, we can claim that even an attacker who makes an error of at most one half wrongly guesses a population frequency with a probability as much as α_ϵ.

A different estimator, which may be biased, can decrease the probability of the incorrect estimation of n_j, but α_ϵ quantifies the deniability of disclosure. For an attacker to claim disclosure, the error probability has to be significantly small. In this sense, α_ϵ is a nominal significance level.

The selection of α_ϵ is more intuitive than that of ϵ. Hence we rewrite eq. (6) as $\epsilon = \log(1 + 2/\sqrt{\alpha_\epsilon})$. For example, when $\alpha_\epsilon = 10\%$, $\epsilon = 1.991$. When $\alpha_\epsilon = 5\%$, $\epsilon = 2.297$. When $\alpha_\epsilon = 1\%$, $\epsilon = 3.045$. These values are comparable to those we have seen in the previous section.

4 Conclusion

Identity protection has been a working device to support the practice of official statistics. It allows the statistical estimation of unknown facts, while a statistical agency can claim that confidentiality is ensured. However, identity protection is more difficult than ever, as we have seen in Section 2.3. Stronger protection such as synthesis is more appreciated.

In this context, fully synthetic data require us to reconsider the meaning of identity protection; we can not accept a logic that fully synthetic data are always publishable because of no presence of real entities. No data should be published when they breach confidentiality.

The present article regards the breach of confidentiality as (identity) disclosure defined in Section 2.5. Identity protection is originally a tool for confidentiality, and a tool itself should not be a goal to pursue. The presence of an entity in data is relevant to identity protection, but it is not necessarily relevant to the breach of confidentiality.

Disclosure has been mathematically defined to be a deterministic situation, which serves legal or institutional requirements. In addition, its uncertainty can be clarified to measure. We now understand that although deterministic masking can prevent disclosure by changing the sample space, universal protection ought to be provided by random masking.

DP is very popular in the assessment of random masking, but not many people seem to understand that DP is just a sufficient condition to bound the accuracy of the estimation of true values of data as shown in Section 3.3. This fact simplifies the decision of a privacy budget, as we have seen in Section 3.5

The remaining issue is the concrete mechanism of random masking that satisfies DP without distorting data too much. Bowen and Liu (2020) compare various ϵ-DP methods by means of simulation studies. Many existing methods are algorithmically defined, and they often employ numerical optimization. Hence the general property of them is difficult to elucidate. Also most of them can not fix a sample size m, which should be consistent with a census table.

Machanavajjhala et al. (2008) propose an exceptional ϵ-DP mechanism: A Dirichlet-multinomial mixture. Its detailed property is widely known, while m is fixed. We note that it has the same support of the multinomial distribution. Such a mechanism can be regarded as (nonsimple) random sampling with replacement, since the multinomial distribution is equivalent to simple random sampling with replacement. Random sampling with replacement can thus be ϵ-DP. This view suggests a direction to improve the Dirichlet-multinomial mixture, which distorts data very much.

We remember that Fienberg (1994) proposes to generate synthetic data by bootstrapping from smoothed raw data; see Section 2.4. Bootstrapping is equivalent to the multinomial distribution as seen in Example 4. Therefore, generating data by nonsimple random sampling with replacement from slightly smoothed raw data can be ϵ-DP; we have to allocate a positive probability to sample points of zero population frequency, though. This type of data much contain characteristics derived from real entities. Simultaneously, DP assures that identity disclosure and attribute disclosure are explicitly deniable. This is the goal of official statistics to protect confidentiality.

The current practice of publishing microdata depends much on subsampling or random sampling without replacement, although this process is not ϵ-DP; see Remark 1. Subsampling, however, is in a broad sense sampling from raw data. Hence random sampling with replacement, which can have DP, should be considered as a method of subsampling.

Many people seem to believe that ϵ-DP microdata must be sampled from a statistical model since microdata including real entities represent "a violation of the core principles of DP" (Bambauer et al., 2013). However, complicated modeling of a synthetic population is not necessary for DP. Just subsampling, nonsimple random sampling with replacement, from smoothed raw data generates differentially private microdata of real entities.

Against subsampling, Dwork (2011) claims "Suppose appearing in a subsample has terrible consequences. Then every time subsampling occurs some individual suffers horribly." However, owing to DP, we can deny terrible consequences or disclosures irrespective of presence in a subsample. It is time to discard the notion of presence.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP18H00835 and JP20H00576.

References

1. Abowd, J.M. and Vilhuber, L. (2008). How protective are synthetic data? Domingo-Ferrer and Saygun (eds) *Privacy in Statistical Databases*. Lecture Notes in Computer Science, **5262**, 239–246, Springer, New York.
2. Aggarwal, C.C. and Yu, P.S. (2004). A Condensation Approach to Privacy Preserving Data Mining. Bertino E. et al. (eds) *Advances in Database Technology - EDBT 2004*. Lecture Notes in Computer Science, **2992**, 183–199, Springer, Berlin.
3. Aggarwal, C.C. and Yu, P.S. (2008). *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, New York.
4. Agrawal, R. and Srikant, R. (2000) Privacy preserving data mining. *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, 439–450.
5. Anderson, M.J. and Seltzer, W. (2009). Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues. *Journal of Privacy and Confidentiality*, **1**, 7–52.
6. Baayen, R.H. (2001). *Word Frequency Distributions*, Kluwer, Dordrecht.
7. Bambauer, J., Muralidhar, K., and Sarathy, R. (2013). Fool's gold: an illustrated critique of differential privacy. Vanderbilt Journal of Entertainment and Technology Law, **16**, 701–755.
8. Barbaro, M. and Zeller, T. (2006). A Face is exposed for AOL searcher no. 4417749, *The New York Times*.

9. Birnbaum, A. (1962). On the foundation of statistical inference. *Journal of the American Statistical Association*, **57**, 269–306.
10. Beckman, R.J., Baggerly, K.A. and McKay, M.D. (1996). Creating synthetic baseline populations, *Transportation Research*, Part A: Policy and Practice, **30**, 415–429.
11. Benedetto, G., Stanley, J.C., and Totty, E. (2018) The Creation and Use of the SIPP Synthetic Beta v7.0, U.S. Census Bureau.
12. Bethlehem, J.G., Keller, W.J., and Pannekoek. J. (1990) Disclosure Control of Microdata, *J. Amer. Statist. Assoc.*, **85**, 38–45.
13. Bishop, Y.M.M., Fienberg, S.E. and Holland. P.W. (1975). *Discrete multivariate analysis: Theory and practice*, MIT Press, Cambridge.
14. Bowen, C.M. and Liu, F. (2020). Comparative study of differentially private data synthesis methods. *Statist. Sci.*, **35**, 280–307.
15. Brand, R. (2002). Microdata Protection through Noise Addition. *Inference Control in Statistical Databases: From Theory to Practice*, Domingo-Ferrer (ed.), Lecture Notes in Computer Science, **2316**, 97–116, Springer, Berlin.
16. Brandt, M., Lenz, R. and Rosemann, M. (2008). Anonymisation of Panel Enterprise Microdata – Survey of a German Project *Privacy in Statistical Databases*, Domingo-Ferrer et al. (eds.), Lecture Notes in Computer Science, **5262**, 139–151, Springer, Berlin.
17. Butz, W. and Torrey, B. (2006) Some Frontiers in Social Science. *Science*, **312**, 1898–1900.
18. Chapman, D.G. and Robbins, H. (1951). Minimum variance estimation without regularity assumptions. *Ann. Math. Statist.*, **22**, 581–586.
19. Chaudhuri, K. and Mishra, N. (2006). When Random Sampling Preserves Privacy. *Proceedings of the 26th Annual International Conference on Advances in Cryptology (CRYPTO 2006)*, 198–213, Springer, Berlin.
20. Clifton, C. and Tassa, T. (2013). On syntactic anonymity and differential privacy. *Transactions on Data Privacy*, **6**, 161–183.
21. Dalenius, T. (1986). Finding a needle in a haystack — or identifying anonymous census records. *Journal of Official Statistics*, **2**, 329–336.
22. Danker, F.K. and El Eman, K. (2013) Practicing differential privacy in health care: A review. *Transactions on Data Privacy*, **5**, 35—67.
23. Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, **11**, 427–444.
24. Deng, M., Wuyts, K., Scandariato, R., Preneel, B. and Joosen, W. (2011). A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering*, **16**, 3–32.
25. Dennis, J.C. (2000). *Privacy and Confidentiality of Health Information*. Jossey-Bass, San Francisco.
26. Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 202–210.
27. Domingo-Ferrer, J. and Tora, V. (2004). *Privacy in Statistical Databases*, Lecture Notes in Computer Science, **3050**, Springer, Berlin, Heidelberg.
28. D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical matching: Theory and practice*, Wiley, Chichester.
29. Doyle, P., Lane, J.I., Theeuwes, J.J.M. and Zayatz, L.V. (2001). *Confidentiality, Disclosure, and Data Access*. Elsevier, Amsterdam.
30. Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Lecture Notes in Statistics, **201**, Springer, New York.
31. Duncan, G.T., Elliot, M. and Salazar-González, J.J. (2011). *Statistical Confidentiality*. Springer, New York.
32. Dwork, C. (2006). Differential privacy. *33rd International Colloquium on Automata, Languages and Programming-ICALP 2006*, Part II, Lecture Notes in Computer Science, **4052**, 1–12, Springer.
33. Dwork, C. (2011). A firm foundation for private data analysis *Communications of the ACM*, **54**, 86–95.
34. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. and Naor, M. (2006a). Our data, ourselves: privacy via distributed noise generation. *Advances in Cryptology - EUROCRYPT 2006*, Vaudenay, S. (ed.), Lecture Notes in Computer Science, **4004**, 486–503, Springer, Berlin.

35. Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. *TCC 2006-Theory of Cryptography Conference*, 265–284.
36. Dwork, C., Smith, A., Steinke, T. and Ulllman, J. (2017). Exposed! A survey of attacks on private data. *Annual Review of Statistics and Its Application*, **4**, 61–84.
37. Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, **7**, 1–26.
38. El Emam, K. and Arbuckle, L. (2013). *Anonymizing Health Data*. O'Reilly, Sebastopol:CA.
39. Erlingsson, U., Pihur, V. and Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. *Proceedings of the 21st ACM Conference on Computer and Communications Security*, ACM, Scottsdale, Arizona.
40. Evett, I., Jackson, G., Lambert, J.A. and McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, **40**, 233–239.
41. Fienberg, S.E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical report, Department of Statistics, Carnegie Mellon University.
42. Fienberg, S.E. (2005). Confidentiality and disclosure limitation. *Encyclopedia of Social Measurement*, Kempf-Leonard (ed.), **1**, 463–469, Elsevier, New York.
43. Fienberg, S.E. and Holland, P.W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, **68**, 683–691.
44. Fienberg, S.E., Makov, U.E. and Steele, R.J. (1998) Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, **14**, 485–502.
45. Fung, B.C.M., Wang, K., Fu, A.W.C. and Yu, P.S. (2010). *Introduction to Privacy-Preserving Data Publishing*. Chapman and Hall/CRC, Boca Raton: FL.
46. Ghosh, A., Roughgarden, T. and Sundararajan, M. (2012). Universally utility-maximizing privacy mechanism. *SIAM Journal of Computing*, **41**, 1673–1693.
47. Giessing, S. (2004). Survey on methods for tabular data protection in ARGUS. *Privacy in Statistical Databases*, Domingo-Ferrer and Torra (eds.), Lecture Notes in Computer Science, **3050**, 1–13, Springer, Berlin.
48. Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, B, **17**, 268–278.
49. Goel, V. (2014). How Facebook sold you krill oil, *The New York Times*.
50. Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.
51. Gottschalk, S. (2004). Microdata disclosure by resampling — Empirical findings for business survey data. *Allgemeines Statistisches Archiv*, **88**, 279–302.
52. Hammersley, J.M. (1950). On estimating restricted parameters. *J. Roy. Statist. Soc.*, Ser. B, **12**, 192–240.
53. Heard, D., Dent, G., Schifeling, T. and Banks, D. (2015) Agent-based models and microsimulation *Annual Review of Statistics and Its Application*, **2**, 259–272.
54. Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.
55. Hoshino, N. (2009). The quasi-multinomial distribution as a tool for disclosure risk assessment. *Journal of Official Statistics*, **25**, 269–291.
56. Hoshino, N. (2016). Evidence based anonymization. *Journal of the Japan Statistical Society*, Series J, **46**, 1–42, (In Japanese.)
57. Hoshino, N. (2018). The control of statistical inference. Talk at Computer Security Symposium 2018, October 24. (In Japanese.)
58. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K. and de Wolf, P.P. (2012). *Statistical Disclosure Control*. Wiley, West Sussex.
59. Inusah, S. and Kozubowski, T.J. (2006). A discrete analogue of the Laplace distribution. *Journal of Statistical Planning and Inference*, **136**, 1090–1102.
60. Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, **31**, 203–222.
61. Jeffreys, H. (1961). *Theory of probability*, 3rd ed., Oxford University Press, Oxford.
62. Kasivisiwanathan, S.P. and Smith, A. (2014). On the Semantics of Differential Privacy: A Bayesian Formulation. *Journal of Privacy and Confidentiality*, **6**, 1–16.
63. Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
64. Khmaladze, E.V. (1987). The statistical analysis of a large number of rare events. *Technical Report Report MS-R8804*, Department of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

65. Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11)*, 193–204, Association for Computing Machinery, New York, NY, USA,
66. Kifer, D. and Machanavajjhala, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, **39**, [a3]. `https://doi.org/10.1145/2514689`
67. Kotz, S., Kozubowski, T. and Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhäuser, Boston.
68. Lee, J. and Clifton, C. (2011). How much is enough? Choosing ϵ for differential privacy. Lai et al. (eds.) *ISC 2011*, Lecture Notes in Computer Science, **7001**, 325–340.
69. Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed., Springer, New York.
70. Li, N., Li, T. and Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. *IEEE 23rd International Conference on Data Engineering (ICDE)*, 106–115.
71. Lindell, Y. and Pinkas, B. (2000). Privacy Preserving Data Mining. *Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology (CRYPTO '00)*, Mihir Bellare (Ed.), 36–54, Springer, London.
72. Little, R. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, **9**, 407–426.
73. Liu, C., He, X., Chanyaswad, T., Wang, S. and Mittal, P. (2019). Investigating statistical privacy frameworks from the perspective of hypothesis testing. *Proceedings on Privacy Enhancing Technologies*, **2019(3)**, 233–254.
74. Lowrance, W.W. (2012). *Privacy, Confidentiality, and Health Research*. Cambridge University Press, New York.
75. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE'08*, 277–286.
76. Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam (2007). ℓ-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, **1(1)**, Article 3.
77. 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.
78. Meiser, S. (2018). Approximate and Probabilistic Differential Privacy Definitions. *IACR Cryptology ePrint Archive*, 2018, 277.
79. Mendes, R. and Vilela, J.P. (2017). Privacy-Preserving Data Mining: Methods, Metrics, and Applications *IEEE Access*, **5**, 10562–10582.
80. Muralidhar, K., Saraty, R. and Li, H. (2016). Secure attribute sharing of linked microdata. *Decision Support Systems*. **81**, 20–29.
81. Nakamura, H. (2017). Microdata access for official statistics in Japan. *Sociological Theory and Methods*, **32**, 310–320. (In Japanese.)
82. National Research Council (2007). *Putting people on the map: Protecting confidentiality with linked social-spatial data*, The National Academies Press, Washington, DC.
83. Neyman, J. and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*. Series A, Containing Papers of a Mathematical or Physical Character, **231**, 289–337.
84. Nin, J. and Herranz, J. (2010). *Privacy and Anonymity in Information Management Systems*. Springer, London.
85. Nissim, K., Raskhodnikova, S. and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. *Proceedings of the Annual ACM Symposium on Theory of Computing*, 75–84.
86. O'Keefe, C.M. (2015). Privacy and Confidentiality in Service Science and Big Data Analytics. *Privacy and Identity Management for the Future Internet in the Age of Globalisation*, Camenisch J., Fischer-Hubner S., Hansen M. (eds), Privacy and Identity 2014. IFIP Advances in Information and Communication Technology, **457**, 54–70, Springer, Cham.
87. Pawitan, Y. (2001). *In All Likelihood*. Clarendon Press, Oxford.
88. Pfitzmann, A. and Hansen, M. (2010). A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. Version 0.34 August 2010, Technical Report, TU Dresden and ULD Kiel. (`http://dud.inf.tu-dresden.de/Anon_Terminology.shtml`)

89. President's Council of Advisors on Science and Technology (2014). *Report to the president: Big data and privacy: A technological perspective*. Executive Office of the President, Washington, DC.
90. Quatember, A. (2015). *Pseudo-Populations*, Springer, Cham.
91. Raab, G.M., Nowok, B. and Dibben, C. (2017). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, **7**, 67–97.
92. Reiter, J.P. (2019). Differential Privacy and Federal Data Releases. *Annual Review of Statistics and Its Application*, **6**, 85–101.
93. Rinott, Y., O'Keefe, C., Shlomo, N., and Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Sciences*,**33**, 358–385.
94. Ritchie, F. (2008). Secure access to confidential microdata: four years of the Virtual Microdata Laboratory. *Economic and Labour Market Review*, **2**, 29–34.
95. Ritchie, F. (2017). The "Five Safes": A framework for planning, designing and evaluating data access solutions. Paper presented at Data for Policy 2017, London, UK.
96. Rocher, L., Hendrickx, J.M. and de Montjoye, Y. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, **10**, 3069.
97. Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
98. Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, **9**, 462–468.
99. Ruggles, S., Fitch, C.A., Magnuson, D.L., and Schroeder, J.P. (2019). Differential Privacy and Census Data: Implications for Social and Economic Research. *AEA Papers and Proceedings*, **109**, 403–408.
100. Shlomo, N. and Skinner, C.J. (2012). Privacy protection from sampling and perturbation in survey microdata. *Journal of Privacy and Confidentiality*, **4**, 155-169.
101. Shlosser, A. (1981). On estimation of the size of the dictionary of a long text on the basis of a sample. *Engineering Cybernetics*, **19**, 97–102.
102. Singer E., Mathiowetz, N.A. and Couper, M.P. (1993) The impact of privacy and confidentiality concerns on survey participation: the case of the 1990 U.S. Ceusus. *Public Opin. Q.*, **57**, 465–482.
103. Singer, E., Van Hoewyk, J. and Neugebauer, R.J. (2003). Attitudes and behavior: the impact of privacy and confidentiality concerns on participation in the 2000 Census. *Public Opin. Q.*, **67**, 368–384.
104. Smith, A. (2008). Efficient, Differentially Private Point Estimators. `https://arxiv.org/abs/0809.4794`
105. Snoke, J., Raab, G., Nowok, B., Dibben, C. and Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society*, Series A, **181**, 663–688.
106. Solove, D.J. (2008). *Understanding Privacy*. Harvard University Press, Cambridge.
107. Solove, D.J. (2013). Privacy self-management and the consent dilemma. *Harvard Law Review*, **126**, 1880–1903.
108. Soria-Comas, J., Domingo-Ferrer, J., Sanchez, D. and Megias, D. (2017). Individual differential privacy: A utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, **12**, 1418–1429.
109. Stewart, K.A. and Segars, A.H. (2002). An empirical examination of the concern for information privacy instrument. *Information Systems Research*, **13**, 36–49.
110. Sweeney, L. (2000). Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh.
111. Sweeney, L. (2002). *k*-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, **10**, 557–570.
112. Tang, J., Korolova, A., Bai, X., Wang, X. and Wang, X. (2017). Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. arXiv:1709.02753 [cs.CR]
113. Templ, M. (2017). *Statistical Disclosure Control for Microdata*. Springer, Cham.
114. Templ, M., Meindl, B., Kowarik, A. and Dupriez, O. (2017). Simulation of synthetic complex data: The `R` package `simPop`, *Journal of Statistical Software*, **79**, 1–38.
115. Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, Addison-Wesley.
116. Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63–69.
117. Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, **66**, 884–888.

118. Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, **105**, 375–389.
119. Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, **111**, Springer, New York.
120. Willenborg, L. and de Waal, T. (2000). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics, **155**, Springer, New York.
121. Wilks, S.S. (1938). The large-sample distribution of the Likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, **9**, 60–62.
122. Zhu, T., Li, G., Zhou, W. and Yu, P.S. (2017). *Differential Privacy and Applications*. Springer, Cham.