Prepared for handbook

Ian M. Schmutte and Lars Vilhuber

2020-08-31

Contents

0.1 Overview of disclosure avoidance methods

The purpose of this handbook is to provide guidance on how to enable broader but ethical and legal access to data. Data providers need to create "safe data" that can be provided to trusted "safe users", within "safe settings", subject to legal and contractual safeguards. Related, but distinct, is the question of how to create "safe outputs" from researchers' findings, before those findings finally make their way into the public through, say, policy briefs or the academic literature. The processes used to create "safe data" and "safe outputs" - manipulations that render data less sensitive and therefore more appropriate for public release - are generally referred to as ||statistical disclosure limitation|| (SDL).[1] In this chapter, we will describe methods traditionally used within the field of SDL, pointing at methods as well as metrics to assess the resultant statistical quality and sensitivity of the data. Newer methods, generally referred to as "formal privacy methods," are described in a separate chapter (Nissim et al NEED REF).

At their core, SDL methods prevent outsiders from learning "too much" about any one record in the data (Dalenius, 1977) by deliberately, and judiciously, adding distortions. Ideally, these distortions maintain the validity of the data for statistical analysis, but strongly reduce the ability to isolate records and infer precise information about individual people, firms, or cases. In general, it is necessary to sacrifice validity in order to prevent disclosure (Abowd & Schmutte, 2015; Goroff, 2015). It is therefore important for data custodians to bear this tradeoff in mind when deciding whether and how to use SDL.

One key challenge for implementing privacy systems lies in choosing the amount, or type, of privacy to provide. Answering this question requires some way to understand the individual and social value of privacy. J. Abowd & Schmutte (2019) discuss the question of optimal privacy protection (see also Hsu et al. (2014) in the specific context of differential privacy). For an illustration, see Spencer & Seeskin (2015), who use a calibration exercise to study the costs, measured in misallocated congressional seats, of reduced accuracy in population census data.

Part of the social value of privacy arises from its relationship to scientific integrity. While the law of information recovery suggests that improved privacy must come at the cost of increased error in published

[1] Other terms sometimes used are "anonymization" or "de-identification," but as this chapter will show, de-identification is a particular method of SDL, and anonymization is a goal, never fully achieved, rather than a method.

statistics, these effects might be mitigated through two distinct channels. First, people may be more truthful in surveys if they believe their data is not at risk, as Couper, Singer, Conrad, & Groves (2008) illustrate. Second, work in computer science and statistics (Dwork et al., 2015, pp. @dwork_fienberg_2018, @cummings_adaptive_2016) suggests a somewhat surprising benefit of differential privacy: protection against overfitting.

There are three factors that a data custodian should bear in mind when deciding whether and how to implement an SDL system in support of making data accessible. First, it is necessary to clarify the specific privacy requirements based on the nature of the underlying data, institutional and policy criteria, and ethical considerations. In addition, the custodian, perhaps in consultation with users, should clarify what sorts of analyses the data will support. Finally, SDL is always to be used as part of a broader system involving access restrictions and other technical barriers. The broader system may allow for less stringent SDL techniques when providing data to researchers in secure environments than would be possible if data were to be released as unrestricted public use data.[2] This implies in particular that we will not provide a recommendation for a "best" method, since no such globally optimal method exists in isolation.

Rather, we provide here an overview of the concepts behind SDL and some of the more widely-used methods. Our hope is to provide a reference point from which data providers and data users can discuss which forms of SDL are appropriate and satisfy the needs of both parties. In particular, we will focus on how common SDL tools affect different types of statistical analysis as well as the kind of confidentiality protections they support, drawing heavily on Abowd & Schmutte (2015). SDL is a broad topic with a vast literature, starting with Fellegi (1972). Naturally, this brief summary is not a replacement for the textbook treatment of SDL in Duncan, Elliot, & Salazar-González (2011). Finally, SDL methods must be implemented and deployed, and we provide pointers to existing "off-the-rack" tools in a variety of platforms (Stata, R, and Python). Readers might also consult other summaries and guides, such as Dupriez & Boyko (2010), World Bank (n.d.), Kopper, Sautmann, & Turitto (2020), and Liu (2020).

0.1.1 Purpose of statistical disclosure limitation methods: Definitions and Context

A clear and precise sense of what constitutes an unauthorized disclosure is a prerequisite to implementing SDL. Are all data items equally sensitive? How much more should one be able to learn about certain classes of people, firms, villages, etc.? Note that even when "trusted researchers" ("safe people") can be sworn to secrecy, the ultimate goal is to publish using information gleaned from the data, and the final audience can never be considered trusted.

The key concepts are privacy and confidentiality. ||Privacy|| can be viewed, in this context, as the right to restrict others' access to personal information, whether through query or through observation (Hirshleifer, 1980). ||Confidentiality|| pertains to data that have already been collected, and describes the principle that the data should not be used in ways that could harm the persons that provided it.

> For example, Ann, who is asked to participate in a study about health behaviors, has a *privacy* right to refuse to answer a question about smoking. If she does answer the question, it would breach *confidentiality* if her response was then used by an insurance company to adjust her premiums (Duncan, Jabine, & Wolf, 1993).

Harris-Kojetin et al. (2005) define "disclosure" as the "inappropriate attribution of information to a data subject, whether an individual or an organization" (Harris-Kojetin et al., 2005, p. 4), and describe three different types of disclosure. An *identity disclosure* is one where it is possible to learn that a particular record or data item belongs to a particular participant (individual or organization). An *attribute disclosure* happens if publication of the data reveals an attribute of a participant. Note that an *identity disclosure* necessarily entails attribute disclosure, but the reverse is not the case.

> In our hypothetical health study, if Ann responds that she is a smoker, an identity disclosure would mean someone can determine which record is hers, and therefore can also learn that she is a smoker – an attribute disclosure. However, an attribute disclosure could also occur if someone

[2] The chapter on IAB provides a good illustration of how various SDL methods are combined with different access methods to provide multiple combinations of analytic validity and risk of disclosure.

> knows that Ann was in the study, they know that Ann lives in a particular zip code, and the data reveal that all participants from that zip code are also smokers. Her full record is not revealed, but confidentiality was breached all the same.

With these concepts in mind, it is necessary to ask whether it is sufficient to prevent blatant "all-or-nothing" identity or attribute disclosures. Usually not, as it may be possible to learn a sensitive attribute with high, but not total, certainty. This is called an *inferential disclosure* (Dalenius, 1977; Duncan & Lambert, 1986).

> Suppose Ann's health insurer knows that Ann is in the data, and that she lives in a particular zip code. If the data have 100 records from that zip code and 99 are smokers, then the insurer has learned Ann's smoking status with imperfect, but high precision.

In addition to deciding what kinds of disclosure can be tolerated and to what extent, in many cases it may also be meaningful to decide which characteristics are and are not sensitive. Smoking behavior may nowadays be regarded as sensitive, but depending on the context, gender might not be. Or, in the case of business data, total sales volume or total payroll are highly sensitive trade secrets. Generally, the county in which the business is located, or the industry in which it operates might not be, but consider a survey of self-employed business people - the location of the business might be the home address, which might be considered highly sensitive. These decisions on what is sensitive affect the implementation of a privacy protection system.

However, additional care must be taken because variables that are not inherently sensitive can still be used to isolate and identify records. Such variables are sometimes referred to as ||*quasi-identifiers*|| and they can be exploited for ||*reidentification*|| attacks. In business data, if the data show that there is only one firm operating in a particular county and sector, then their presence inherently leads to identity disclosure. Many of the traditional approaches to SDL operate in large part by preventing re-identification.[3] Garfinkel (2015) discusses techniques for de-identifying data and the many ways in which modern computing tools and a data-rich environment may render effective de-identification impossible, reinforcing the growing need for formal privacy models like differential privacy.

In the United States, 62% of individuals are aware (and possibly resigned) that government and private companies collect data on them, and seem to believe that there is little benefit to them of such collection: 81% think so when companies do the data collection, and 66% when the government does so (Auxier et al., 2019).

There is a large and robust literature on the value of privacy in economics. However, that literature is generally focused on the value to individuals of being able to conceal private information, like a health condition, from a firm or prospective employer. The challenge to the firm is to design mechanisms, like pricing strategies, that encourage people to disclose private information. For an overview of ideas in this literature, we recommend reading Stigler (1980), Posner (1981), and Hirshleifer (1980). Varian (2002) and Acquisti, Taylor, & Wagman (2016) both provide comprehensive surveys at different points in the development of this literature.

SDL methods may be required for legal and ethical reasons. ||Institutional Review Boards|| require that individual's well-being be protected (see [chapter on IRB]). Legal mandates may intersect with ethical concerns, or prescribe certain (minimal) criteria. Thus, the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA) (U.S. Department of Health & Human Services, n.d.) has precise definitions of variables that need to be removed in order to comply with the law's mandate of deidentification (Department of Health and Human Services, 2012). The European ||General Data Protection Regulation|| (GDPR) came into effect in 2018, and has defined both broadly the way researchers can access data, and more narrowly the requirements for disclosure limitation [Cohen & Nissim (2020);greene_adjusting_2019;molnar-gabor_germany_2018]. Similar laws are emerging around the world, and will define both minimal requirements and limits of SDL and other access controls. The ||California Consumer Privacy Act|| (CCPA) (Marini, Kateifides, & Bates, 2018) and the Brazilian ||Lei Geral de Proteção de Dados|| (LGDP) (Black, Ramos, & Biscardi, 2020) came into effect in 2020, and India is currently considering such a law (Panakal, 2019).

[3] Thus the occasional reference to methods as *deidentification* or *anonymization*, though these terms can sometimes be misleading in regard to what they can actually achieve.

0.1.2 Methods

There are many different SDL methods, and the decision of which to use depends on what needs to be protected, how their use will affect approved analyses, and their technical properties. At a very high level, we can think of an SDL system as a mechanism that takes the raw confidential data, D, as an input and produces a modified dataset, $\tilde{D}$. The researcher then conducts their analysis with the modified $\tilde{D}$. Ideally, the researcher can do their analysis as planned, but the risk of disclosure in $\tilde{D}$ is reduced.

Researchers generally need to consider all of the design features that went into producing the data used for an analysis. Most already do so in the context of surveys, where design measures are incorporated into the analysis, often directly in software packages. Some of these adjustments may already take into account various SDL techniques. Traditional survey design adjustments can take into account sampling. Some forms of coarsening may already be amenable to adjustment using various clustering techniques, such as Moulton (1986) (see Cameron & Miller (2015)).

More generally, the inclusion of edits to the data done in service of disclosure limitation is less well supported, and less well integrated into standard research methods. Abowd & Schmutte (2015) argue that the analysis of SDL-laden data is inherently compromised because the details of the SDL protections cannot be disclosed. If they cannot be disclosed, their consequences for inference are unknowable, and, as they show, may be substantial. Regression models, regression discontinuity designs, and instrumental variables models are, generally speaking, affected when SDL is present. The exact nature of any bias or inconsistency will depend on whether SDL was applied to explanatory variables, dependent variables, instruments, or all of the above. Furthermore, it is not always the case that SDL induces an attenuating bias.

With these goals in mind, following Abowd & Schmutte (2015) we distinguish between *ignorable* and *non-ignorable* SDL systems. Briefly, SDL is *ignorable* for a particular analysis if the analysis can be performed on the modified data, $\tilde{D}$ as though it were the true data. In a non-ignorable analysis, the result differs in some material way when $\tilde{D}$ is substituted for D. When the SDL method is *known*, then it may be possible for the researcher to perform an *SDL-aware* analysis that corrects for non-ignorability. However, SDL methods are almost never ignorable. Non-ignorable SDL may be discoverable if and only if the analysis of the published data can be probabilistically corrected for the data alterations introduced by the SDL. Many SDL methods have this latter property.

We briefly outline several of the methods most commonly used within national statistical offices. For interested readers, Harris-Kojetin et al. (2005) describe how SDL systems are implemented in the U.S. statistical system,[4] while Dupriez & Boyko (2010) offers a more multinational perspective.

0.1.2.1 De-identification In general, it is good practice to remove any variables from the data that are not needed for data processing or analysis, and that could be considered direct identifiers. What constitutes "direct identifiers" may differ on the context, but generally comprises any variable that might directly link to confidential information: names, account or identifier numbers, and sometimes exact birth dates or exact geo-identifiers.[5] HIPAA defines sixteen identifiers that must be removed in order to comply with the law. It may be necessary to preserve identifiers through parts of the data processing or analysis if they are key variables needed for record linking. For instance, names may be used to link records between surveys and administrative data, or precise geographic coordinates may be needed to compute distances as part of the analysis. Hence, the specific application will restrict the set of applicable data protection systems.

0.1.2.2 Suppression Suppression is perhaps the most common form of SDL, and one of the oldest (Fellegi, 1972). In their most basic form, suppression rules work as follows:

1. Model the sensitivity of a particular data item, table cell, or observation ("disclosure risk")
2. Do not allow the release of data items that have excessive disclosure risk (primary suppression)
3. Do not allow the release of other data from which the sensitive item can be calculated (complementary suppression)

[4] As of the writing of this chapter in August 2020, WP22 is being revised and updated, but has not yet been published.
[5] See guidance in World Bank (n.d.) and Kopper et al. (2020) .

Suppression rules can be applied to microdata, in which case the sensitive observations are removed from the microdata, or to tabular data, where the relevant cells are suppressed.

In the case of business microdata, a firm that is unique in its county and industry might be flagged as having high disclosure risk and eliminated from the data. Another less damaging possibility is that just its sensitive attributes are suppressed, so a researcher would still know that there was a firm operating in that industry and location; just not what its other attributes were. For tabular data, the principle is the same. Continuing with the business application, suppose there is one large firm and several smaller competitors in a given industry and location. If the cell is published, it might be possible for its local competitors to learn the receipts of the dominant firm to a high degree of precision.

Cell suppression rules based on this sort of reasoning are called p-percent rules, where p describes the precision with which the largest firm's information can be learned. A conservative estimate of this occurs when the largest firm's value is *(1-p)%* of the cell's value.
A variant of this rule takes into account prior precision q (the "pq percent rule"). Another rule is known as the *n,k* rule: a cell is suppressed if n or fewer entities contribute k percent or more of the cell's value. These rules are frequently applied to statistics produced by national statistical agencies (Harris-Kojetin et al., 2005). Simpler rules based entirely on cell counts are also encountered, for instance in the Health and Retirement Study (Health and Retirement Study, n.d.). Tables produced using HRS confidential geo-coded data are only allowed to display values when the cell contains three or more records (five for marginal cells).

If a cell in a contingency table is suppressed based on any one of these rules, it could be backed out by using the information in the table margins and the fact that table cells need to sum up to their margins. Some data providers therefore require that additional cells are suppressed to ensure this sort of reverse engineering is not possible. Figuring out how to choose these *complementary suppressions* in an efficient manner is a non-trivial challenge.

In general, cell suppression is not an ignorable form of SDL. It remains popular because it is easy to explain and does not affect the un-suppressed cells.

Data suppression is clearly non-ignorable, and is quite difficult to correct for suppression in an SDL-aware analysis.[6] The features of the data that lead to suppression are often related to the underlying phenomenon of interest. Chetty & Friedman (2019) provide a clear illustration. They publish neighborhood-level summaries of intergenerational mobility based on tax records linked to Census data. The underlying microdata are highly sensitive, and to protect privacy they used a variant of a differentially privacy model. They show that if they had instead used a cell suppression rule, the published data would be misleading with respect to the relationship between neighborhood poverty and teen pregnancy because both variables are associated with neighborhood population. Hence, the missingness induced by cell suppression is not ignorable.

Suppression can also be applied to model-based statistics. For instance, after having run a regression, coefficients that correspond to cells with fewer than n cases may be suppressed. This most often occurs when using dichotomous variables (dummy variables), which represent conditional means for particular subgroups.

> In a regression, a researcher includes a set of dummies for interacting occupation and location. When cross-tabulating occupation and location, many cells have less than 5 observations contributing to the coefficient. The data provider requires that these be suppressed.

0.1.2.3 Coarsening Coarsening takes detailed attributes that can serve as quasi-identifiers and collapses them into a smaller number of categories. Computer scientists call this "generalizing", and it is also sometimes referred to as "masking". Coarsening can be applied to quasi-identifiers, to prevent re-identification, or to attributes, to prevent accurate attribute inference. When applied to quasi-identifiers, the concern is that an outsider could use detailed quasi-identifiers to single-out a particular record and learn who it belonged to. By coarsening quasi-identifiers, the set of matching records is increased, raising uncertainty about any re-identified individual's true identity. In principle, all variables can serve as quasi-identifiers, and the concept

[6] One approach is to replace suppressed cells with imputed values, and then treat the data as multiply-imputed.

of *k-anonymity* introduced by Sweeney (2002) is a useful framework for thinking about how to implement coarsening and other microdata SDL. We discuss k-anonymity in the section on disclosure risk.

Coarsening is common in microdata releases. As a general rule of thumb, it may make sense to consider coarsening variables with heavy tails (earnings, payroll), residuals (truncate range, suppress labels of range). In public-use microdata from the American Community Survey, geographic areas are coarsened until all such areas represent at least 60,000 individuals (REFERENCE). In many data sources, characteristics like age and income, are reported in bins even when the raw data are more detailed. Income topcoding is another common type of coarsening, in which incomes above a certain threshold are replaced with some topcoded value (e.g., $200,000 in the Current Population Survey). When releasing model-based estimates, rounding can satisfy statistical best practice – not releasing numbers beyond their statistical precision – as well as disclosure avoidance principles – by preventing inferences that could be too precise about specific records in the data.

Whether coarsening is ignorable or not depends on the analysis to be performed. Consider the case in which incomes are topcoded above the 95th percentile. This form of SDL is ignorable with respect to estimating the 90th percentile of the income distribution (and all other quantiles below the 95th). However, coarsening age is not ignorable if the goal is to conduct an analysis of behavior of individuals right around some age or date-of-birth cutoff. Coarsening rules should therefore bear in mind the intended analysis for the data and may be usefully paired with restricted-access protocols that allow trusted researchers access to the more detailed data. See Burkhauser, Feng, Jenkins, & Larrimore (2011) for an example of the impact of top-coding on estimates of earnings inequality.

0.1.2.4 Swapping The premise behind the technique of *swapping* is similar to suppression. Again, each record is assigned a level of disclosure risk. Then, any high-risk record is matched to a less risky record on a set of key variables, and all of the other non-key attributes are swapped. The result is a dataset that preserves the distribution among all the key variables used for matching. If the original purpose of the data was to publish cross-tabulations of the matching variables, swapping can produce microdata that are consistent with those tabulations. This approach is more commonly used in censuses and surveys of people or households, and rarely used with establishment data.

> For example, consider our hypothetical health study again, and now suppose we know Ann's zip code, gender, race, ethnicity, age, smoking behavior, and the size of her household. Ann's record might be classified as high risk if, say, she has a very large household relative to the rest of the other respondents who are also from her zip code. If the data are used to publish, say, summaries of smoking behavior by age, race, and gender, then Ann's record would be matched to another record with the same age, race, gender and smoking behavior, and the values of the household size and zip code attributes would be swapped.

Swapping is ignorable for analyses that only depend on the matching variables, since the relationships among them will be preserved. However, swapping distorts relationships among the other variables, and between the matching variables and the other variables. In the example above, the swapping would be non-ignorable in the context of a study of how smoking behavior various across zip codes. In general, statistical agencies are not willing to publish detailed information about how swapping is implemented since that information could be used to reverse-engineer some of the swaps, undoing the protection. Hence, SDL-aware analysis may not be possible, and inference validity negatively affected.

0.1.2.5 Sampling Sampling is the original SDL technique. Rather than the full confidential microdata, publishing a sample inherently limits the certainty with which an attackers can re-identify records. While sampling can provide a formal privacy guarantee, in modern, detailed surveys, sampling will not in general prevent re-identification. In combination with other tools, like coarsening, sampling may be particularly appealing because, while it is non-ignorable, researchers can adjust their analysis for the sampling using familiar methods. Sampling is often used in conjunction with other methods, including with formally private methods, to amplify the protection provided.

0.1.2.6 Noise infusion Noise infusion can refer to an array of related methods, all of which involve distorting data with randomly distributed noise. There is a key distinction between methods where the microdata are infused with noise (||input noise infusion||), versus methods where noise is added to functions or aggregates of the data before publication (||output noise infusion||).

Noise infusion was developed as a substitute for cell suppression as an approach to protecting tabular summaries of business data. Originally proposed by Evans, Zayatz, & Slanta (1998), the basic approach assigns each microdata unit (a business establishment) a multiplicative noise factor drawn from a symmetric distribution (e.g., centered on one), and multiplies sensitive (or all) characteristics by that factor. Tabular summaries can then be made from the distorted characteristics. As cell sizes increase, the distortions applied to each unit average out. Thus, while small cells may be quite distorted and thus protected, large cells usually have little distortion. Most cells no longer need to be suppressed. These approaches are used in the U.S. Census Bureau's Quarterly Workforce Indicators (Abowd et al., 2012, 2009) and County Business Patterns, with a truncated distribution. When the noise distribution is unbounded, for instance Gaussian, noise infusion may be differentially private, see [chapter on DP].

Noise infusion has the advantage that it mostly eliminates the need to suppress sensitive records or cells, allowing more information to be revealed from the confidential data while maintaining certain confidentiality protections. Noise infusion also generally preserves the means and covariances among variables. However, it will always inflate estimated variances and can lead to bias in estimates of statistical models, and in particular regression coefficients. Hence, noise infusion is, in general, not ignorable. If the details of the noise distribution can be made available to researchers, then it is possible to correct analysis for noise infusion. However, information about the noise distribution can also help an attacker reverse engineer the protections.

0.1.2.7 Synthetic data and multiple imputation Synthetic data generation and multiple imputation are closely related. In fact, one particular variant of synthetic data as SDL – partially synthetic data – is also known as "suppress and impute" (Little, 1993). Sensitive values for some or all records are replaced by (multiple) imputations. More generally, fully synthetic data (Rubin, 1993) replaces all values with draws from a posterior predictive distribution, estimated given the confidential data. For an overview, see Raghunathan, Reiter, & Rubin (2003), Little, Liu, & Raghunathan (2004), and Drechsler (2011).

Synthetic data have been used in the Federal Reserve Board's Survey of Consumer Finances to protect sensitive income values (Kennickell, 1998), and in the American Community Survey microdata to protect data from group quarters (such as prisons and university residences; see Hawala & Rodriguez (2009)). The U.S. Census Bureau's LODES data, included in the OnTheMap application, uses synthetic household data (Machanavajjhala, Kifer, Abowd, Gehrke, & Vilhuber, 2008). Synthetic data can be used in conjunction with validation servers: researchers use the synthetic data to create complex model-based estimation, then submit their analysis to a remote server with access to the confidential data for validation of the results. Such a mechanism has been used by the U.S. Census Bureau in collaboration with Cornell University for confidential business microdata (Kinney et al., 2011) and for survey data combined with administrative data (Abowd, Stinson, & Benedetto, 2006). The term is sometimes used as well for "test" data for remote submission systems, which typically makes no claims as to the validity - it is simply constructed to replicate the data schema of the confidential data, to test statistical code.

0.1.3 Metrics

The design of an SDL system depends on determinations about what constitutes an acceptable level of disclosure risk, balanced with the proposed uses of the data. There are many different ways to describe and measure disclosure risk. These share in common a sense of how unique a record or combination of attributes is in the data, which corresponds intuitively to a sense of how easy it would be to single-out that record and re-identify the respondent, perhaps aided by a linked dataset. Likewise, there are many different ways to assess whether the released data are suitable, or fit, for their intended use. These quality measures are often based on how closely the released data match the true data on certain statistical summaries, and it will be important for researchers and data custodians to agree on what are the most relevant summaries.

0.1.3.1 Disclosure Risk Early definitions of disclosure risk were based on rules and guidelines derived from institutional knowledge, assessment of summary measures, and reidentification experiments (Harris-Kojetin et al. (2005)). Statisticians have subsequently developed more formal models to measure risk of re-identification for specific types of publication and with particular threat models. For instance, Shlomo & Skinner (2010) model reidentification risk in survey microdata when an attacker is matching on certain categorical variables.

Recently, computer scientists and statisticians have introduced more general concepts of disclosure risk and data privacy. Latanya Sweeny proposed the concept of k-anonymity (Sweeney, 2002) which defines disclosure risk in terms of the number of records that share the same combination of attributes. If a single record is uniquely identified by some combination of attributes, disclosure risk is high. Sweeny says that a dataset can be called k-anonymous if for all feasible combinations of attributes, at least k records have that combination. Intuitively, increases in k reduce the risk that observations can be singled out by linking other datasets that contain the same attributes. The concept of k-anonymity can provide some guidance when thinking about how to implement the SDL systems described above. For example, if records are uniquely identified by age, race, and gender, then one might collapse age into brackets until there are at least $k > 1$ records for each such combination.

However, k-anonymity does not protect against attribute disclosure. If all k observations with the same combination of attributes also share the same sensitive attribute, say smoking behavior, then the published data do not fully prevent disclosure of smoking behavior. Recognizing this, Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam (2007) introduce the concept of ℓ-diversity. The idea is that whenever a group of records are identical on some set of variables, there must be a certain amount of heterogeneity in important sensitive traits. If a certain group of records match on a set of quasi-identifiers and also all share the same smoking status, then to achieve ℓ-diversity, we might alter the reported smoking behavior of some fraction (ℓ) of the records - a form of noise infusion.

0.1.3.2 Data Quality When the released data or output is tabular (histograms, cross-tabulations) or is a limited set of population or model parameters (means, coefficients), a set of distance-based metrics (so-called "ℓ_p distance" metrics) can be used to compare the quality of the perturbed data. Note that this is a specific metric, as it is limited to those statistics taken into account - the data quality may be very poor in non-measured attributes! For $p = 1$, the ℓ_1 distance is the sum of absolute differences between the confidential and perturbed data. For $p = 2$, the l_2 distance is the sum of squared differences between the two datasets (normalized by n the number of observations, it is the Mean Squared Error, MSE). In settings where it is important to measure data quality over an entire distribution, the Kullbach-Leibler (KL) divergence measure can also be used. The KL-divergence is related to the concept of entropy from information theory and, loosely, measures the amount of surprise associated with seeing an observation drawn from one distribution when you expected them to come from another distribution. Other metrics are based on propensity scores (Snoke, Raab, Nowok, Dibben, & Slavkovic, 2018; Woo, Reiter, Oganian, & Karr, 2009). More specific measures will often compare specific analysis output, a task that is quite difficult to conduct in general. Reiter, Oganian, & Karr (2009) propose to summarize the difference between regression coefficients when analyses can be run on both confidential and protected data,in the context of verification servers.

0.1.4 Tools

Multiple institutions provide guidance to researchers who wish to apply SDL techniques. We point in particular to a useful checklist by ICPSR (ICPSR, 2020) for useful practical guidance.

Because the particular requirements of a given SDL system are often unique, they are frequently implemented using custom programming. Nevertheless, a few tools are of more general applicability. The listing below is almost certainly incomplete, but should provide practitioners with a starting point in applying SDL for data publication.

Statistics Netherlands maintains the ARGUS software for SDL (Hundepool & Willenborg, 1998), including -ARGUS to protect tabular data (De Wolf, 2018), and -ARGUS for protecting microdata (Hundepool & Ramaswamy, 2018). The software appears to be widely used in statistical agencies in Europe. An open-source R, sdcMicro implements a full suite of tools needed to apply SDL, from computation of risk measures, including k-anonymity and l-diversity, to implementation of SDL methods and the computation of data quality measures (Templ, Kowarik, & Meindl, 2015 ; Templ, Meindl, & Kowarik, 2020).

Simpler tools, focusing on removing direct identifiers, can be found at J-PAL (for Stata (stata_PII_scan) (J-PAL, 2020b) and R (PII-scan (J-PAL, 2020a)) and Innovations for Poverty Action (for Python or Windows PII_detection (Innovations for Poverty Action, 2020)).

A number of R packages facilitate generation of synthetic data. Raab, Nowok, & Dibben (2016) and Nowok, Raab, & Dibben (2016) provide a flexible and up-to-date packages with methods for generating synthetic microdata. Templ et al. (2019) can also generate synthetic populations from aggregate data, which can be useful for testing SDL systems on non-sensitive data. In some cases, one might also consider using general-purpose software for multiple imputation for data synthesis.[7]

0.1.5 Conclusion

Disclosure avoidance attempts to strike a balance between usability of the data (in a particular environment and context) and the ability of others to infer identity and attributes of the people and institutions represented in the data. This chapter provided an overview of the various techniques traditionally used to modify the data to achieve that goal. The techniques range from the simple to complex, with varying degrees of confidentiality protection.

0.1.6 About the Authors

Ian M. Schmutte is Associate Professor in the Department of Economics at the University of Georgia. Schmutte is currently working with the U.S. Census Bureau on new methods for protecting confidential data. His research has appeared in the American Economic Review, Journal of Labor Economics, Journal of Human Resources, Journal of Business and Economic Statistics, and the Brookings Papers on Economic Activity.

Lars Vilhuber is the Executive Director of the Labor Dynamics Institute at Cornell University. He has worked for many years with the Research and Methodology Directorate at the U.S. Census Bureau on a variety of projects, including implementing disclosure avoidance techniques. He is a member of governing or scientific committees of secure data access center in Canada (CRDCN and France (CASD, and a member of the American Statistical Association‘s Committee on Privacy and Confidentiality. He is the inaugural Data Editor for the American Economic Association, and the Managing (Executive) Editor of the Journal of Privacy and Confidentiality. He is the Co-Chair, Innovations in Data and Experiments for Action (IDEA) Initiative at J-PAL.

0.1.7 Acknowledgements

This chapter draws on Abowd et al. (2019) and the INFO7470 class at Cornell University Abowd & Vilhuber (2016) . We gratefully acknowledge the support of Alfred P. Sloan Foundation Grant G-2015-13903 and NSF Grant SES-1131848 for the earlier work.

0.1.8 Disclaimer

The views expressed in this paper are those of the authors and not those of the U.S. Census Bureau or other sponsors.

0.1.9 References

Abowd, J. M., Gittings, R. K. K., McKinney, K. L., Stephens, B., Vilhuber, L., & Woodcock, S. D. (2012). *Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures for Related Time Series* (No. 12-13). https://doi.org/10.2139/ssrn.2159800

[7] See “Multiple imputation in Stata” or the mice package in R (Buuren & Groothuis-Oudshoorn, 2011).

Abowd, J. M., Schmutte, I., Sexton, W., & Vilhuber, L. (2019). *Introductory Readings in Formal Privacy for Economists* (Document No. 2662639). https://doi.org/10.5281/zenodo.2662639

Abowd, J. M., Stephens, B. E., Vilhuber, L., Andersson, F., McKinney, K. L., Roemer, M., & Woodcock, S. D. (2009). The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. In T. Dunne and J B Jensen and M J Roberts (Ed.), *Producer Dynamics: New Evidence from Micro Data*. University of Chicago Press.

Abowd, J. M., Stinson, M., & Benedetto, G. (2006). *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project* (No. 1813/43929). Retrieved from U.S. Census Bureau website: http://hdl.handle.net/1813/43929

Abowd, J. M., & Vilhuber, L. (2016). *Session 12: Statistical Tools: Methods of Confidentiality Protection* (Presentation No. 45060). Retrieved from Labor Dynamics Institute, Cornell University website: https://hdl.handle.net/1813/45060

Abowd, J., & Schmutte, I. (2019). An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*, *109*(1), 171–202. https://doi.org/10.1257/aer.20170627

Abowd, J., & Schmutte, I. M. (2015). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, 221–267. https://doi.org/10.1353/eca.2016.0004

Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*, *54*(2), 442–492. https://doi.org/10.1257/jel.54.2.442

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019). *Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information*. Retrieved from Pew Research Center website: https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Black, K., Ramos, G. A., & Biscardi, G. (2020). 6 Months Until Brazil's LGPD Takes Effect – Are You Ready? Retrieved from https://www.natlawreview.com/article/6-months-until-brazil-s-lgpd-takes-effect-are-you-ready

Burkhauser, R. V., Feng, S., Jenkins, S. P., & Larrimore, J. (2011). Estimating trends in US income inequality using the Current Population Survey: The importance of controlling for censoring. *The Journal of Economic Inequality*, *9*(3), 393–415. https://doi.org/10.1007/s10888-010-9131-6

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from https://www.jstatsoft.org/v45/i03/

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Chetty, R., & Friedman, J. N. (2019). A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples. *Journal of Privacy and Confidentiality*, *9*(2). https://doi.org/10.29012/jpc.716

Cohen, A., & Nissim, K. (2020). Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8344–8352. https://doi.org/10.1073/pnas.1914598117

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *Journal of Official Statistics*, *24*(2), 255. Retrieved from http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/risk-of-disclosure-perceptions-of-risk-and-concerns-about-privacy-and-confidentiality-as-factors-in-survey-participation.pdf

Cummings, R., Ligett, K., Nissim, K., Roth, A., & Wu, Z. S. (2016). Adaptive Learning with Robust Generalization Guarantees. *CoRR*, *abs/1602.07726*. Retrieved from http://arxiv.org/abs/1602.07726

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, *15*, 429–444. https://doi.org/10.1145/320613.320616

Department of Health and Human Services. (2012). Methods for De-identification of PHI [Text]. Retrieved from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

De Wolf, P.-P. (2018). *T-ARGUS*. Retrieved from http://research.cbs.nl/casc/tau.htm

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. https://doi.org/10.1007/978-1-4614-0326-5

Duncan, G., & Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, *81*(393), 10–18. https://doi.org/10.1080/01621459.1986.10478229

Duncan, G. T., Elliot, M., & Salazar-González, J.-J. (2011). *Statistical confidentiality: Principles and practice*. https://doi.org/10.1111/j.1751-5823.2012.00196_11.x

Duncan, G. T., Jabine, T. B., & Wolf, V. A. de (Eds.). (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. https://doi.org/10.17226/2122

Dupriez, O., & Boyko, E. (2010). *Dissemination of Microdata Files - Principles, Procedures and Practices* (Working Paper No. 005). Retrieved from The World Bank website: http://ihsn.org/dissemination-of-microdata-files

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). Generalization in Adaptive Data Analysis and Holdout Reuse. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 2341–2349). Retrieved from http://papers.nips.cc/paper/5993-generalization-in-adaptive-data-analysis-and-holdout-reuse.pdf

Dwork, C., & Ullman, J. (2018). The Fienberg Problem: How to Allow Human Interactive Data Analysis in the Age of Differential Privacy. *Journal of Privacy and Confidentiality*, *8*(1). https://doi.org/10.29012/jpc.687

Evans, T., Zayatz, L., & Slanta, J. (1998). Using Noise for Disclosure Limitation of Establishment Tabular Data. *Journal of Official Statistics*, *14*(4), 537–551.

Fellegi, I. P. (1972). On the Question of Statistical Confidentiality. *Journal of the American Statistical Association*, *67*(337), 7–18. https://doi.org/10.2307/2284695

Garfinkel, S. (2015). *De-Identification of Personal Information* (Internal Report No. 8053). https://doi.org/10.6028/nist.ir.8053

Goroff, D. L. (2015). Balancing privacy versus accuracy in research protocols. *Science*, *347*(6221), 479–480. https://doi.org/10.1126/science.aaa3483

Harris-Kojetin, B. A., Alvey, W. L., Carlson, L., Cohen, S. B., Cohen, S. H., Cox, L. H., … Groves, R. (2005). *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology* [Research Report]. Retrieved from U.S. Federal Committee on Statistical Methodology website: https://nces.ed.gov/FCSM/pdf/spwp22.pdf

Hawala, S., & Rodriguez, R. (2009). *Disclosure avoidance for group quarters in the American Community Survey: Details of the synthetic data method* [Presentation]. Retrieved from https://ecommons.cornell.edu/handle/1813/47676

Health and Retirement Study. (n.d.). *Disclosure Limitation Review*. Retrieved from https://hrs.isr.umich.edu/data-products/restricted-data/disclosure-limitation-review

Hirshleifer, J. (1980). Privacy: Its origin, function, and future. *The Journal of Legal Studies*, *9*(4), 649–664. https://doi.org/10.1086/467659

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., & Roth, A. (2014). Differential Privacy: An Economic Method for Choosing Epsilon. *2014 IEEE 27th Computer Security Foundations Symposium*, 398–410. https://doi.org/10.1109/CSF.2014.35

Hundepool, A., & Ramaswamy, R. (2018). *M-ARGUS*. Retrieved from http://research.cbs.nl/casc/mu.htm

Hundepool, A., & Willenborg, L. (1998). ARGUS, Software Packages for Statistical Disclosure Control. In R. Payne & P. Green (Eds.), *COMPSTAT* (pp. 341–345). https://doi.org/10.1007/978-3-662-01131-7_45

ICPSR. (2020). *Disclosure Risk Worksheet* (Document No. 156095). Retrieved from University of Michigan website: http://hdl.handle.net/2027.42/156095

Innovations for Poverty Action. (2020). *PovertyAction/PII_detection.* Retrieved from https://github.com/PovertyAction/PII_detection

J-PAL. (2020a). *J-PAL/PII-Scan.* Retrieved from https://github.com/J-PAL/PII-Scan

J-PAL. (2020b). *J-PAL/stata_PII_scan.* Retrieved from https://github.com/J-PAL/stata_PII_scan

Kennickell, A. B. (1998). Multiple imputation in the Survey of Consumer Finances. *Proceedings of the Section on Survey Research.* Retrieved from https://www.federalreserve.%20gov/econresdata/scf/files/impute98.pdf

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, *79*(3), 362–384. https://doi.org/10.1111/j.1751-5823.2011.00153.x

Kopper, S., Sautmann, A., & Turitto, J. (2020). *J-PAL Guide to de-identifying data* (p. 12). Retrieved from J-PAL Global website: https://www.povertyactionlab.org/sites/default/files/research-resources/J-PAL-guide-to-deidentifying-data.pdf

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *9*(2), 407–426. Retrieved from http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf

Little, R. J. A., Liu, F., & Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 141–152). Retrieved from https://doi.org/10.1002/0470090456.ch13

Liu, F. (2020). A Statistical Overview on Data Privacy. *arXiv:2007.00765 [Cs, Stat].* Retrieved from http://arxiv.org/abs/2007.00765

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008). Privacy: Theory meets practice on the map. *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, 277–286. https://doi.org/10.1109/ICDE.2008.4497436

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1). https://doi.org/10.1145/1217299.1217302

Marini, A., Kateifides, A., & Bates, J. (2018). *Comparing privacy laws: GDPR v. CCPA*. Retrieved from Future of Privacy Forum website: https://fpf.org/wp-content/uploads/2018/11/GDPR_CCPA_Comparison-Guide.pdf

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, *32*(3), 385–397. https://doi.org/10.1016/0304-4076(86)90021-7

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, *74*(1), 1–26. https://doi.org/10.18637/jss.v074.i11

Panakal, D. D. (2019). India's Proposed Privacy Law Allows Government Access and Some Data Localization. Retrieved from https://www.natlawreview.com/article/india-s-proposed-privacy-law-allows-government-access-and-some-data-localization

Posner, R. A. (1981). The Economics of Privacy. *The American Economic Review*, *71*(2), 405–409. Retrieved from https://www.jstor.org/stable/1815754

Raab, G. M., Nowok, B., & Dibben, C. (2016). Practical Data Synthesis for Large Samples. *Journal of Privacy and Confidentiality*, *7*(3), 67–97. https://doi.org/10.29012/jpc.v7i3.407

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, *19*(1). Retrieved from http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/multiple-imputation-for-statistical-disclosure-limitation.pdf

Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, *53*(4), 1475–1482. https://doi.org/10.1016/j.csda.2008.10.006

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468. Retrieved from http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

Shlomo, N., & Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Annals of Applied Statistics*, *4*(3), 1291–1310. https://doi.org/10.1214/09-AOAS317

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(3), 663–688. https://doi.org/10.1111/rssa.12358

Spencer, B. D., & Seeskin, Z. H. (2015). Effects of Census Accuracy on Apportionment of Congress and Allocations of Federal Funds. *JSM Proceedings, Government Statistics Section*, 3061–3075. Retrieved from https://www.ipr.northwestern.edu/publications/papers/2015/ipr-wp-15-05.html

Stigler, G. J. (1980). An Introduction to Privacy in Economics and Politics. *Journal of Legal Studies*, *9*(4), 623–644. https://doi.org/10.2307/724174

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 571–588. https://doi.org/10.1142/s021848850200165x

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software*, *67*(4). https://doi.org/10.18637/jss.v067.i04

Templ, M., Kowarik, A., Meindl, B., Alfons, A., Ribatet, M., & Gussenbauer, J. (2019). *simPop: Simulation of Synthetic Populations for Survey Data Considering Auxiliary Information.* Retrieved from https://CRAN.R-project.org/package=simPop

Templ, M., Meindl, B., & Kowarik, A. (2020). *sdcMicro: Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation.* Retrieved from https://CRAN.R-project.org/package=sdcMicro

U.S. Department of Health & Human Services. (n.d.). *Health Information Privacy.* Retrieved from https://www.hhs.gov/hipaa/index.html

Varian, H. R. (2002). Economic Aspects of Personal Privacy. In W. H. Lehr & L. M. Pupillo (Eds.), *Cyber Policy and Economics in an Internet Age* (pp. 127–137). https://doi.org/10.1007/978-1-4757-3575-8_9

Woo, M., Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *Privacy and Confidentiality*, *1*(1), 111–124. Retrieved from http://repository.cmu.edu/cgi/viewcontent.cgi?article=1006&context=jpc

World Bank. (n.d.). *DIME Wiki: De-identification.* Retrieved from https://dimewiki.worldbank.org/wiki/De-identification