CHAPTER 5

Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods

Ian M. Schmutte (University of Georgia)

Lars Vilhuber (Cornell University)

5.1   Introduction

The purpose of this Handbook is to provide guidance on how to en-
able broader but ethical and legal access to data. Within the Five Safes
framework (Desai, Ritchie and Welpton, 2016), data providers need
to create *safe data* that can be provided to trusted *safe people* for use
within *safe settings* (chapter 2), subject to legal and contractual safe-
guards (chapter 3). Related, but distinct, is the question of how to
create *safe outputs* from researchers' findings before those findings fi-
nally make their way into the public through, for example, policy briefs
or the academic literature. The processes used to create safe data and
safe outputs (manipulations that render data less sensitive and there-
fore more appropriate for public release) are generally referred to as

statistical disclosure limitation (SDL).[1] This chapter will describe techniques 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 approaches, generally referred to as *formal privacy methods*, are described in chapter 6.

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 (Goroff, 2015; Abowd and Schmutte, 2015). It is therefore important for data custodians to bear this trade-off 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. Abowd and 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 and 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 statistics, these effects might be mitigated through two distinct channels. First, people may be more truthful in surveys if they believe their data are not at risk (Couper et al., 2008). Second, work in computer science and statistics (Dwork et al., 2015; Dwork and Ullman, 2018; Cummings et al., 2016) suggests a somewhat surprising benefit of differential privacy: protection against overfitting.

---

[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.

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 often part of a broader system to protect sensitive data that can also involve 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 that the chapter will not provide a recommendation for a "best" method, since no such globally optimal method exists in isolation.

Rather, this chapter provides an overview of the concepts and more widely used methods of SDL. Relative to other primers that cover similar material, this text focuses more closely on the advantages and disadvantages of various methods from the perspective of data users. This chapter can serve as a reference that data providers and data users can employ to discuss which forms of SDL are appropriate and will satisfy the needs of both parties. In particular, there is a focus on how common SDL tools affect different types of statistical analysis as well as the kind of confidentiality protections these tools support, drawing heavily on Abowd and 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 and Salazar-González (2011). Finally, SDL methods must be implemented and deployed, and the chapter provides pointers to existing off-the-rack tools in a variety of platforms (Python, R, and Stata). Readers might also consult other summaries and guides, such as Dupriez and Boyko (2010), World Bank (n.d.), Kopper, Sautmann and Turitto (2020), and Liu (2020).

---

[2]Chapter 7 on the RDC-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.

5.2   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.[3]

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 their information.

> 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 and de 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). They proceed to 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

---

[3]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).

of a participant. Note that an *identity disclosure* necessarily entails *attribute disclosure*, but the reverse is not the case.

> In the 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 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 was 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 and 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. 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 the business operates might not be sensitive, 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.[4]

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 *re-identification* 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 attempting to prevent re-identification.[5] 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.

SDL methods may be required for legal and ethical reasons. Institutional review boards (IRBs) require that individual's well-being be protected (see chapter 4 on IRBs). Legal mandates may intersectwith ethical concerns, or prescribe certain (minimal) criteria. Thus, the US 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 de-identification (Department of Health and Human Services, 2012). The European Union General Data Protection Regulation (GDPR) came into effect in 2018 and has defined both the way researchers can access data and the requirements for disclosure limitation (Cohen and Nissim, 2020; Greene et al., 2019; Molnár-Gábor, 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 and Bates, 2018) and the Brazilian Lei Geral de Proteção de Dados (LGPD) (Black, Ramos and Biscardi, 2020) came into effect in 2020,

---

[4]There is a large and robust literature in economics on the value of privacy. For an overview of ideas in this literature, we recommend Varian (2002) and Acquisti, Taylor and Wagman (2016).

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

and India is currently considering such a law (Panakal, 2019).

Finally, note that there is a parallel concept of non-statistical disclosure limitation that is a complementary part of secure data dissemination. This applies to the metadata—like codebooks, data descriptions, and other summary information—that can leak potentially sensitive information. For example, data documentation might reveal that only certain geographic areas were included in a particular collection, information that could be used as an element in a re-identification attack. While typically not considered quantitative disclosure avoidance, some of the same concepts described here can apply to such metadata as well. For instance, removing mention of the collection area from the documentation is akin to suppression, while only revealing broad regions of data collection is akin to coarsening.

5.3 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 high level, think of an SDL system as a mechanism that takes the raw confidential data, D, as inputs and produces a modified data set, $\tilde{D}$. 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 consider sampling. Some forms of coarsening may already be amenable to adjustment using various clustering techniques (Moulton, 1986; Cameron and Miller, 2015).

More generally, the inclusion of edits to the data done in service of disclosure limitation is less well supported by, and less well integrated in, standard research methods. Abowd and Schmutte (2015) argue that

the analyses of SDL-laden data are inherently compromised because the details of the SDL protections cannot be disclosed. If the details cannot be disclosed, the consequences for inference are unknowable and, as they show, may be substantial. Regression models, regression discontinuity designs, and instrumental variables models are generally 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 and Schmutte (2015), this chapter distinguishes 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 generally not ignorable except in certain specific applications.

The chapter briefly outlines several of the methods most commonly used within national statistical offices. For interested readers, Harris-Kojetin et al. (2005)[6] describe how SDL systems are implemented in the US statistical system, while Dupriez and Boyko (2010) offers a more multinational perspective.

5.3.1 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. This is often referred to as de-identification. 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

---

[6]As of the writing of this chapter in August 2020, WP22 is being revised and updated, but has not yet been published.

sometimes exact birth dates or exact geo-identifiers.[7] HIPPA 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. In field experiments, the identities of treatment and control units may need to be merged with an administrative data set. It is also sometimes necessary to use direct identifiers to link records between surveys and administrative data, or precise geographic coordinates may be needed to compute distances as part of the analysis. If possible, the data provider should facilitate record linking while the data are secure and before they are shared with the research team.

5.3.2 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).

Suppression rules can be applied to microdata: 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 the sensitive attributes are suppressed, so a researcher would still know that there was a firm operating in that industry and location but not the other attributes. For tabular data, the principle is the same. Continuing with the business application, suppose there is one large firm

---

[7]See guidance in World Bank (n.d.) and Kopper, Sautmann and Turitto (2020).

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's original value could be backed out by using the information in the table margins and the understanding 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 it is quite difficult to correct for suppression in an SDL-aware analysis.[8] The features of the data that lead to suppression are often related to the underlying phenomenon of interest. Chetty and Friedman (2019) provide a clear

---

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

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 the researchers used a variant of a differentially privacy model. Chetty and Friedman 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 five observations contributing to the coefficient. The data provider requires that these be suppressed.

5.3.3   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 to whom it belonged. 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 of *k-anonymity* introduced by Sweeney (2002) is a useful framework for thinking about how to implement

coarsening and other microdata SDL. *K-anonymity* is discussed in section 5.4.1.

Coarsening is common in microdata releases. Generally, 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 100,000 individuals (U.S. Census Bureau, 2011). In many data sources, characteristics like age and income, are reported in bins even when the raw data are more detailed. Topcoding is a common type of coarsening in which variables, such as incomes above a certain threshold, are replaced with some topcoded value (e.g., US$200,000 in the Current Population Survey). When releasing model-based estimates, rounding (another form of coarsening) 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 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 et al. (2011) for an example of the impact of topcoding on estimates of earnings inequality.

5.3.4 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 data set 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.

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 varies 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.

> For example, consider the hypothetical health study again, and now suppose the known factors are 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, for example, 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 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.

5.3.5 Sampling

Sampling is the original SDL technique. Rather than the full confidential microdata, publishing a sample inherently limits the certainty

with which 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.

5.3.6 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 and 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 US Census Bureau's Quarterly Workforce Indicators (Abowd et al., 2009, 2012) 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 6 on differential privacy).

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 confiden-

tiality 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 generally 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.

5.3.7 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 and Rubin (2003), Little, Liu and 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 US Census Bureau's American Community Survey microdata to protect data from group quarters such as prisons and university residences (Hawala and Rodriguez, 2009). The US Census Bureau's LODES data, included in the OnTheMap application, uses synthetic household data (Machanavajjhala et al., 2008). Synthetic data can be used in conjunction with validation servers: researchers use the synthetic data to create complex model-based estimation and 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 US 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 and 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.

5.3.8 Examples of SDL Methods

Table 5.1 shows how the various methods can be combined, drawing on examples both from this Handbook as well as from other frequently used data sources.

5.4 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. A commonality these systems share is the ability to determine the uniqueness of a record, or combination of attributes in the data, that then intuitively predicts the ease with which a record could be distinguished to re-identify the respondent (perhaps aided by a linked data set). 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.

5.4.1 Disclosure Risk

Early definitions of disclosure risk were based on rules and guidelines derived from institutional knowledge, assessment of summary measures, and re-identification 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 and Skinner (2010)

Table 5.1: Summary of SDL methods

	In this Hand-book	Remov-al of Direct Identi-fiers	Remov-al of Quasi-Identi-fiers	Supp-ress-ion	Coars-ening	Swapp-ing	Sampl-ing	Noise Infus-ion	Synth-etic Data
IAB On-Site Access	✔	✔	✘	✔	✔	✘	✔	✘	✘
IAB Scientific Use Files	✔	✔	✔	✔	✔	✘	✔	✘	✘
OLDA	✔	✔	◼	✔	✔	✘	✘	✘	✘
NB-IRDT	✔	✔	—	✔	✔	✘	✘	✘	✘
PCRI	✔	✔	—	✘	—	✘	✘	✘	✘
Aurora Public-Use File	✔	✔	◼	✔	✔	✘	✘	✘	✘
Stanford-SFUSD	✔	◼	✔	✔	✘	✘	✘	✘	✘
City of Cape Town	✔	◼	—	—	✔	—	✘	✘	—
DIME (World Bank)	✔	◼	✔	—	—	—	—	✘	—
Survey of Con-sumer Fi-nances	✘	✔	—	✔	✔	✘	✔	✘	✔
American Com-mu-nity Survey	✘	✔	—	✘	✔	✔	✔	✘	◼
Quarterly Work-force Indica-tors	✘	✔	✘	✔	✔	✘	✘	✔	✘

Notes: ✔ = Yes, ◼ = Partially, ✘ = No, — = No Info

model re-identification 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 Sweeney 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. Sweeney says that a data set 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 data sets 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, for example, smoking behavior, then the published data do not fully prevent disclosure of smoking behavior. Recognizing this, Machanavajjhala et al. (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 matches on a set of quasi-identifiers and also all share the same smoking status, then to achieve ℓ-diversity, one might alter the reported smoking behavior of some fraction (ℓ) of the records—a form of noise infusion.

5.4.2 Data Quality

When the released data or output are tabular (histograms, cross-tabulations) or are 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 data sets (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 Kullback-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 one expected them to come from another distribution. Other metrics are based on propensity scores (Woo et al., 2009; Snoke et al., 2018). More specific measures will often compare specific analysis output, a task that is quite difficult to conduct in general. Reiter, Oganian and 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.

5.5 Tools

For data providers faced with the need to start providing safe data for use by external researchers, a growing number of software packages are available that implement the methods described in this chapter. The Inter-university Consortium for Political and Social Research (ICPSR) has a checklist that may be of use in early development of an SDL system (ICPSR, 2020). The listing of tools below is incomplete but will provide practitioners with a place to start. A fully developed SDL system will have unique requirements and may require custom programming. Nevertheless, many tools are useful across a wide range of applications.

Statistics Netherlands maintains the ARGUS software for SDL (Hundepool and Willenborg, 1998), including τ-ARGUS to protect tabular

data (De Wolf, 2018), and μ-ARGUS for protecting microdata (Hunde-pool and Ramaswamy, 2018). The software appears to be widely used in statistical agencies in Europe. An open-source R package, `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 and Meindl, 2015; Templ, Meindl and Kowarik, 2020).

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

A number of R packages facilitate generation of synthetic data. Raab, Nowok and Dibben (2016) and Nowok, Raab and Dibben (2016) provide `synthpop`, a flexible and up-to-date package with methods for generating synthetic microdata. The R package `simPop` (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.[9]

Many of the methods described in this chapter are technical and require statistical and programming expertise. If that expertise is not already available among staff, some institutions provide guidance to researchers who wish to apply SDL techniques.

5.6 Conclusion

There is now a greater demand for all kinds of data. More than ever before, scholars and analysts have the tools to use data to better understand the economy and society and to inform policy. Alongside these advances, data custodians find themselves under pressure to make databases available to outsiders. However, the pressure to make data

---

[9]See "Multiple imputation in Stata" or the `mice` package in R (Buuren and Groothuis-Oudshoorn, 2011).

available is not always accompanied by the resources, tools, or expertise needed to do so safely.

The same advances driving these new demands have a darker side. Computing power together with the availability of detailed outside data make it easier than ever for attackers to exploit improperly protected data. Therefore, when making data available for research, agency stewards must take great care to also protect the subjects in the data. This chapter provides an overview of techniques traditionally used to modify the data to achieve that goal. There is a legitimate concern that some of the methods discussed here cannot protect against all possible attacks made possible with modern computing power. Those concerns animate the discussion of formal methods that yield provable privacy guarantees elsewhere in this Handbook.

About the Authors

Ian M. Schmutte is Associate Professor in the Department of Economics at the University of Georgia. Ian is currently working with the US 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 & Economic Statistics*, and the *Brookings Papers on Economic Activity*.

Lars Vilhuber is the Executive Director of the Labor Dynamics Institute at Cornell University, and a faculty member in Cornell University's Economics Department. He is also the American Economic Association's Data Editor. Lars is a Co-Chair of IDEA. His research interests relate to the dynamics of the labor market. He also has extensive experience in the application of privacy-preserving publication and access to restricted data. He is chair of the scientific committee of the French restricted-access system CASD, member of the governing board of the Canadian Research Data Centre Network (CRDCN), and incoming chair of the American Statistical Association's Committee on Privacy and Confidentiality. Lars has an undergraduate degree in Economics from Universität Bonn and a Ph.D. in Economics from Université de Montréal.

Disclaimer

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

Acknowledgements

References in Chapter 5

Abowd, John M., and Ian M. Schmutte. 2015. "Economic analysis and statistical disclosure limitation." *Brookings Papers on Economic Activity*, 221–267. https://doi.org/10.1353/eca.2016.0004.

Abowd, John M., and Ian M. Schmutte. 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, John M., and Lars Vilhuber. 2016. "Session 12: Statistical Tools: Methods of Confidentiality Protection." Labor Dynamics Institute, Cornell University Presentation 45060. https://hdl.handle.net/1813/45060.

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock. 2009. "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators." In *Producer Dynamics: New Evidence from Micro Data*. , ed. Timothy Dunne, J. Bradford Jensen and Mark J. Roberts. University of Chicago Press. https://www.nber.org/chapters/c0485.

Abowd, John M., Ian M. Schmutte, William Sexton, and Lars Vilhuber. 2019. "Introductory Readings in Formal Privacy for Economists." Labor Dynamics Institute, Cornell University Document 2662639, https://doi.org/10.5281/zenodo.2662639.

Abowd, John M., Martha Stinson, and Gary Benedetto. 2006. "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project." U.S. Census Bureau 1813/43929. http://hdl.handle.net/1813/43929.

Abowd, John M., R. Kaj Gittings, Kevin L. McKinney, Bryce Stephens, Lars Vilhuber, and Simon D. Woodcock. 2012. "Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures for Related Time Series." U.S. Census Bureau, Center for Economic Studies 12-13, https://doi.org/10.2139/ssrn.2159800.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman. 2016. "The Economics of Privacy." *Journal of Economic Literature*, 54(2): 442–492. https://doi.org/10.1257/jel.54.2.442.

Auxier, Brooke, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. 2019. "Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information." Pew Research Center. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/ (accessed 2020-07-17).

Black, Kate, Gretchen A. Ramos, and Giovanni Biscardi. 2020. "6 Months Until Brazil's LGPD Takes Effect – Are You Ready?" *The National Law Review*. https://www.natlawreview.com/article/6-months-until-brazil-s-lgpd-takes-effect-are-you-ready (accessed 2020-08-08).

Burkhauser, Richard V., Shuaizhang Feng, Stephen P. Jenkins, and Jeff Larrimore. 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, Stef van, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 45(3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Cameron, A. Colin, and Douglas L. Miller. 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, Raj, and John N. Friedman. 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, Aloni, and Kobbi Nissim. 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, Mick P., Eleanor Singer, Frederick G. Conrad, and Robert M. Groves. 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. 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, Rachel, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. 2016. "Adaptive Learning with Robust Generalization Guarantees." *CoRR*, abs/1602.07726. http://arxiv.org/abs/1602.07726.

Dalenius, Tore. 1977. "Towards a methodology for statistical disclosure control." *Statistik Tidskrift*, 15: 429–444.

Department of Health and Human Services. 2012. "Methods for De-identification of PHI." *HHS.gov*. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (accessed 2020-08-26).

Desai, Tanvi, Felix Ritchie, and Richard Welpton. 2016. "Five Safes: Designing data access for research." https://uwe-repository.worktribe.com/output/914745 (accessed 2020-01-30).

De Wolf, Peter-Paul. 2018. "τ-ARGUS." [software]. http://research.cbs.nl/casc/tau.htm (accessed 2020-08-17).

Drechsler, Jörg. 2011. *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics*, New York:Springer-Verlag. https://doi.org/10.1007/978-1-4614-0326-5.

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

Duncan, George T., Mark Elliot, and Juan-José Salazar-González. 2011. *Statistical confidentiality: principles and practice. Statistics for Social and Behavioral Sciences*, New York:Springer-Verlag. https://doi.org/10.1111/j.1751-5823.2012.00196_11.x.

Duncan, George T., Thomas B. Jabine, and Virginia A. de Wolf, ed. 1993. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* National Academies Press. https://doi.org/10.17226/2122.

Dupriez, Olivier, and Ernie Boyko. 2010. "Dissemination of Microdata Files - Princi-

ples, Procedures and Practices." The World Bank Working Paper 005. http://ihsn.o rg/dissemination-of-microdata-files (accessed 2019-11-15).

Dwork, Cynthia, and Jonathan Ullman. 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.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. 2015. "Generalization in Adaptive Data Analysis and Holdout Reuse." In *Advances in Neural Information Processing Systems 28*. , ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, 2341–2349. http://papers.nips.cc/paper/5 993-generalization-in-adaptive-data-analysis-and-holdout-reuse.pdf.

Evans, Timothy, Laura Zayatz, and John Slanta. 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, Simson. 2015. "De-Identification of Personal Information." National Institute of Standards and Technology Internal Report 8053, https://doi.org/10.6028/ni st.ir.8053.

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

Greene, Travis, Galit Shmueli, Soumya Ray, and Jan Fell. 2019. "Adjusting to the GDPR: The Impact on Data Scientists and Behavioral Researchers." *Big Data*, 7(3): 140–162. https://doi.org/10.1089/big.2018.0176.

Harris-Kojetin, Brian A., Wendy L. Alvey, Lynda Carlson, Steven B. Cohen, Steve H. Cohen, Lawrence H. Cox, Robert E. Fay, Ronald Fecso, Dennis Fixler, Gerald Gates, Barry Graubard, William Iwig, Arthur Kennickell, Nancy J. Kirkendall, Susan Schechter, Rolf R. Schmitt, Marilyn Seastrom, Monroe G. Sirken, Nancy L. Spruill, Clyde Tucker, Alan R. Tupek, G. David Williamson, and Robert Groves. 2005. "Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology." U.S. Federal Committee on Statistical Methodology Research Report. https://nces.ed.gov/FCSM/pdf/spwp22.pdf (accessed 2020-12-15).

Hawala, Sam, and Rolando Rodriguez. 2009. "Disclosure avoidance for group quarters in the American Community Survey: Details of the synthetic data method." http://hdl.handle.net/1813/47676 (accessed 2020-08-17).

Health and Retirement Study. n.d.. "Disclosure Limitation Review." University of Michigan. https://hrs.isr.umich.edu/data-products/restricted-data/disclosure-limit ation-review (accessed 2020-08-09).

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

Hsu, Justin, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C. Pierce, and Aaron Roth. 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, Anco, and Leon Willenborg. 1998. "ARGUS, Software Packages for Sta-

tistical Disclosure Control." 341–345. Heidelberg:Physica-Verlag HD. https://doi.or
g/10.1007/978-3-662-01131-7_45.

Hundepool, Anco, and Ramya Ramaswamy. 2018. "μ-ARGUS." [software]. http:
//research.cbs.nl/casc/mu.htm (accessed 2020-08-17).

ICPSR. 2020. "Disclosure Risk Worksheet." University of Michigan Document 156095.
http://hdl.handle.net/2027.42/156095 (accessed 2020-08-21).

Innovations for Poverty Action. 2020. "PovertyAction/PII_detection." https://github
.com/PovertyAction/PII_detection (accessed 2020-08-30).

J-PAL. 2020*a*. "J-PAL/PII-Scan." https://github.com/J-PAL/PII-Scan (accessed 2020-
08-30).

J-PAL. 2020*b*. "J-PAL/stata_PII_scan." https://github.com/J-PAL/stata_PII_scan (ac-
cessed 2020-08-30).

Kennickell, Arthur B. 1998. "Multiple imputation in the Survey of Consumer Fi-
nances." https://www.federalreserve.gov/econresdata/scf/files/impute98.pdf
(accessed 2020-10-05).

**Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S.
Jarmin, and John M. Abowd.** 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, Sarah, Anja Sautmann, and James Turitto. 2020. "J-PAL Guide to de-
identifying data." J-PAL Global. https://www.povertyactionlab.org/sites/default
/files/research-resources/J-PAL-guide-to-deidentifying-data.pdf (accessed 2020-
08-08).

Little, Roderick JA. 1993. "Statistical analysis of masked data." *Journal of Official
Statistics*, 9(2): 407–426. http://www.scb.se/contentassets/ca21efb41fee47d293b
bee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

Little, Roderick J. A., Fang Liu, and Trivellore E. Raghunathan. 2004. "Statisti-
cal disclosure techniques based on multiple imputation." In *Applied Bayesian model-
ing and causal inference from incomplete-data perspectives*. , ed. Andrew Gelman and
Xiao Li Meng, 141–152. Wiley. https://doi.org/10.1002/0470090456.ch13.

Liu, Fang. 2020. "A Statistical Overview on Data Privacy." *arXiv:2007.00765 [cs, stat]*.
http://arxiv.org/abs/2007.00765 (accessed 2020-08-31).

**Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrish-
nan Venkitasubramaniam.** 2007. "L-diversity: privacy beyond k-anonymity." *ACM
Transactions on Knowledge Discovery from Data*, 1(1). https://doi.org/10.1145/1217
299.1217302.

**Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and
Lars Vilhuber.** 2008. "Privacy: theory meets practice on the map." 277–286. https:
//doi.org/10.1109/ICDE.2008.4497436.

Marini, Alice, Alexis Kateifides, and Joel Bates. 2018. "Comparing privacy laws:
GDPR v. CCPA." Future of Privacy Forum. https://fpf.org/wp-content/uploads/201
8/11/GDPR_CCPA_Comparison-Guide.pdf (accessed 2020-06-20).

Molnár-Gábor, Fruzsina. 2018. "Germany: a fair balance between scientific freedom
and data subjects' rights?" *Human Genetics*, 137(8): 619–626. https://doi.org/10.1

007/s00439-018-1912-1.

Moulton, Brent 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, Beata, Gillian M. Raab, and Chris Dibben. 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, Dominic Dhil. 2019. "India's Proposed Privacy Law Allows Government Access and Some Data Localization." *The National Law Review*. https://www.natlawreview.com/article/india-s-proposed-privacy-law-allows-government-access-and-some-data-localization (accessed 2020-06-20).

Raab, Gillian M., Beata Nowok, and Chris Dibben. 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, Trivellore E., Jerry P. Reiter, and Donald B Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics*, 19(1). http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/multiple-imputation-for-statistical-disclosure-limitation.pdf.

Reiter, Jerome P., Anna Oganian, and Alan F. Karr. 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, Donald B. 1993. "Discussion: Statistical disclosure limitation." *Journal of Official Statistics*, 9(2): 461–468. http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf.

Shlomo, Natalie, and Chris Skinner. 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, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 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, Bruce David, and Zachary H. Seeskin. 2015. "Effects of Census Accuracy on Apportionment of Congress and Allocations of Federal Funds." *JSM Proceedings, Government Statistics Section*, 3061–3075. https://www.ipr.northwestern.edu/our-work/working-papers/2015/ipr-wp-15-05.html.

Sweeney, Latanya. 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, Matthias, Alexander Kowarik, and Bernhard Meindl. 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, Matthias, Alexander Kowarik, Bernhard Meindl, Andreas Alfons, Mathieu Ribatet, and Johannes Gussenbauer. 2019. "simPop: Simulation of Synthetic Pop-

ulations for Survey Data Considering Auxiliary Information." https://CRAN.R-proje ct.org/package=simPop (accessed 2020-08-17).

Templ, Matthias, Bernhard Meindl, and Alexander Kowarik. 2020. "sdcMicro: Sta- tistical Disclosure Control Methods for Anonymization of Data and Risk Estimation." https://CRAN.R-project.org/package=sdcMicro (accessed 2020-08-17).

U.S. Census Bureau. 2011. "Final Public Use Microdata Area (PUMA) Criteria and Guidelines for the 2010 Census and the American Community Survey." https://ww w2.census.gov/geo/pdfs/reference/puma/2010_puma_guidelines.pdf.

U.S. Department of Health & Human Services. n.d.. "Health Information Privacy." https://www.hhs.gov/hipaa/index.html (accessed 2020-06-23).

Varian, Hal R. 2002. "Economic Aspects of Personal Privacy." In *Cyber Policy and Economics in an Internet Age*. , ed. William H. Lehr and Lorenzo M. Pupillo, 127–137. Boston, MA:Springer US. https://doi.org/10.1007/978-1-4757-3575-8_9.

Woo, Mi-Ja, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. 2009. "Global Mea- sures of Data Utility for Microdata Masked for Disclosure Limitation." *Privacy and Confidentiality*, 1(1): 111–124. https://doi.org/10.29012/jpc.v1i1.568.

World Bank. n.d.. "DIME Wiki: De-identification." https://dimewiki.worldbank.org /wiki/De-identification (accessed 2020-08-08).