TRANSACTIONS ON DATA PRIVACY 17 (2024) 55–88

RWN: A Novel Neighborhood-Based Method for Statistical Disclosure Control

Noah Perry*, **Norman Matloff****, **Patrick Tendick*****

*Department of Statistics, University of California, Davis

**Department of Computer Science, University of California, Davis. Corresponding author.

***TEKsystems, Inc.

E-mail: nsmatloff@ucdavis.edu

Received 19 August 2022; received in revised form 27 February 2024; accepted 17 May 2024

Abstract. A novel variation of the data swapping approach to statistical disclosure control is presented, aimed particularly at preservation of multivariate relations in the original dataset for proper statistical analysis. A theorem is proved in support of the method, and extensive empirical investigation is reported.

Keywords. statistical disclosure control; data swapping; k-nearest neighbors; multivariate relations

1 Introduction

The field of statistical disclosure control (SDC)—maintaining privacy of individual records in a dataset while retaining the statistical utility of the data—has long been the subject of arcane technical analysis, conducted mainly by statisticians. The advent of *differential privacy* (DP) methods in 2006 [15] brought computer scientists into the field, and the arcane nature of the SDC field changed with the well-publicized adoption of DP by the United States Census Bureau in 2017 [1].

Different SDC tools may be appropriate for different databases in different settings, not just in terms of numeric degree of protection afforded by a tool, but also in terms of usability, interpretability and transparency for end users [36] [22]. Here we develop new methodology that we believe database curators will find useful in a variety of settings.

Our proposed method, Randomization Within Neighborhoods (RWN), to be presented in Section 4, is inspired by data swapping, a classic approach to SDC. However, RWN differs from previous data swapping approaches, in that it exploits a certain statistical independence property, to be described in detail in Section 7.

Our context here will be that in which the database curator modifies the original microdata just once, and releases the result to the public. This is the setting of most "traditional" SDC techniques.

A key aspect will be the ability of an SDC method to preserve, to the degree possible, multivariate statistical relations. In essence, any SDC method is subject to attenuation or other distortion of such relations. We will refer to one of the goals as Multivariate Relations Attenuation Resistance (MRAR).

This paper is organized as follows. We set the stage with some motivating examples in Section 2. We then give a brief overview of SDC methods in Section 3. The RWN method is presented in

Section 4. This is followed by a discussion of considerations in SDC specific to databases used for statistical analysis in Section 5, and a comparison of RWN to other methods in this regard in Section 6. The underlying theory of RWN is given in Section 7. Tuning parameter selection is covered in Section 8. Our empirical investigation is presented in Section 9, and computational issues are discussed in Section 10.

2 Some Motivating Examples

Data privacy breaches are sometimes due to careless protection of passwords and the like, but frequently arise because the intruder possesses key identifying information about one or more specific records in the database. There are myriad different types and examples of such breaches. In order to motivate our presentation, we cite three here.

A. Intruder has knowledge of certain small cells

Consider for instance the example given in [2]:

> [Say]...in Athens, Georgia, there [is] only one male household head with 10 children...

By submitting a query for the mean income of all such people, knowing there is only one, an intruder can thus illicitly obtain the income of this person. A similar example involving two neighbors who know each other's attributes as the only Hispanics in the block is given in [21]. Indeed, this problem has been known since the early years of SDC; see for instance [8]'s example of the sole female professor in a given university department.

B. Intruder knows the identity of some extreme record

For instance, in an employee database, the intruder may know that a certain worker has the highest salary. In a hospital database, the intruder might know, say, the identity of the patient whose current stay is the longest, and so on. Such settings may lead to re-identification of the record of interest.

This is a setting of frequent concern in SDC applications. As noted in [16], "...outliers may be precisely those people for whom privacy is most important." The **sdcMicro** package's **dRiskRMD()** function, which computes robust Mahalonobis distances, is aimed at this situation, with the documentation noting that

> ...it accounts for the "outlyingness" of each observations. This is a quite natural approach since outliers do have a higher risk of re-identification and therefore these outliers should have larger disclosure risk intervals as observations in the center of the data cloud.

C. Intruder simply wishes to know whether a given entity (person, firm etc.) is represented in the dataset

For example, with a cancer patient dataset, re-identification of a record reveals that the given person has the disease.

3 SDC Methods

Good surveys of SDC methods are in [14] [25] [5]. Our RWN approach is a type of *data swapping*. To motivate RWN, let us review the main SDC classes:

- Cell suppression

  This concerns settings such as the Georgian householder we saw in Section 2. Here any query concerning a very small number of database records is denied.

  As noted, this kind of situation, i.e. queries involving small cells, has been the subject of examples in many SDC papers over the years. It is central to the SDC field, yet the problem has been long known to be highly mathematically challenging and computationally demanding [19].

  An obvious solution is to have the database refuse to provide an answer to such queries. What may be less obvious, though, is that the database must also refuse to supply answers for *complementary* queries. The intruder may, for example, query the total income of all households with 10 children, then query this quantity for the female-headed households of that size, then simply subtract to illicitly obtain the desired information. Thus at least one of these queries must be denied.

  This kind of attack must also be guarded against if there are only two records satisfying a query, rather than just one, as a more complex attack involving several queries and subtractions would give the intruder some private quantity. To solve this problem in general requires sophisticated mathematical tools [19], and a very serious additional concern is that it removes much information from the database of interest to legitimate users.

- Noise addition

  Another standard SDC approach is *noise addition*, in which random noise is added to achieve privacy [26] [28] [42] [41]. Differential privacy, which is technically a privacy *criterion* rather than a privacy *method*, is also typically implemented via some kind of noise addition [10].

- Data swapping

  In the past, the Census Bureau has used data swapping for SDC [19] [14] [18], and it has been proposed in numerous variants over the years, such as [6] [9] [18] [31] [33] [38]. For the purposes of discussion here, we will define data swapping broadly, treating any method that moves or duplicates data as belonging to the "data swapping" class, including *rank swapping*, *microaggregation*[1] and RWN, though we will sometimes distinguish among such methods.

  A set of *key variables* is defined that may render certain individual records in the data vulnerable to disclosure. In the basic form, some records, especially those deemed most at risk to disclosure, will have the values of their key variables swapped with those in other records, say drawn from the same geographic region.

  Data swapping methods are typically applied one variable at a time, thus creating the concern of attenuation of multivariate relations. This aspect will be seen below to comprise a major impetus for our RWN method.

[1] For the sake of convenient exposition, we will treat microaggregation as being in this category but will point out differences at some points.

4 RWN: Randomization Within Neighborhoods

Our method works as follows. For each record in the data, we define a neighborhood using either a given radius or a given number of nearest neighbors. Then, for each record r we randomly choose a subset of the variables to perturb. For each such variable, we replace its original value by its counterpart in a randomly chosen record in the neighborhood of r. A key point is that a different random neighbor record may be used for each of the variables to be perturbed in r.

4.1 Formal Description

Let $W = (w_{ij}), i = 1, ..., n, j = 1, ..., p$ denote our original data on n individuals and p variables.

Choose the neighborhood radius $\epsilon > 0$, number of nearest neighbors k, and modification probability q.

Then form the released/perturbed data $W' = (w'_{ij}), i = 1, ..., n, j = 1, ..., p$ as follows:

For $i = 1, ...n$:

1. Consider record i in the database:

$$r_i = (w_{i1}, ..., w_{ip}) \tag{1}$$

2. Find the set S_i of records within the neighborhood of r_i other than r_i itself. Each neighborhood is defined to be either the k-nearest neighbors of r_i or the set of neighbors within ϵ distance of r_i, whichever set is larger. For the distance computations, categorical variables are dummy/one-hot encoded, and numeric variables are scaled to the $[0, 1]$ interval.
3. For $j = 1, ..., p$:
   With probability $1-q$, leave w_{ij} unmodified, but with probability q, modify it. For variables j that are chosen to be modified, we replace w_{ij} with the value in variable j of a random record in the neighborhood S_i. As noted, there may be a different such random record for each j (records sampled from S_i with replacement). This results in a perturbed data point w'_{ij}. For unmodified variables, $w'_{ij} = w_{ij}$.
4. Store the released, modified version of r_i as

$$r'_i = (w'_{i1}, ..., w'_{ip}) \tag{2}$$

A key point is that in Step 3, the p actions here are *taken independently of each other*. In other words, the process acts as if the p variables in the data are statistically independent of each other. This would at first seem to violate our goal of MRAR, but it is all resolved in the theorem in Section 7 below.

We call this technique Randomization Within Neighborhoods (RWN).

4.2 Comparison to Rank Swapping

Rank swapping [31] does data swapping using ranks rather than data values, in order to facilitate dealing with discrete variables. The standard version of rank swapping is implemented in **sdcMicro**

through the **rankSwap()** function [40]. Although rank swapping has performed well in some empirical evaluations comparing SDC methods [12], numerous variants of the method have been proposed to address some of its limitations and privacy vulnerabilities. RWN addresses many of these issues.

The rank swapping algorithm involves a swapping range, which is analogous to RWN's neigborhoods. Standard rank swapping uses swapping ranges of uniform size, which can be exploited by an intruder by using a form of record linkage specially designed for rank swapped data. This vulnerability was the motivation for two new variants of the method: *rank swapping p-distribution* and *rank swapping p-buckets* [35]. By contrast, as seen in Figure 1, RWN's neighborhoods can be of varying sizes, depending on the choice of tuning parameters.

RWN places a premium on MRAR. Distances are computed based on all variables in a record instead of one variable at a time, so the records within a given neighborhood can be "similar" in many different ways.[2] Rank swapping does achieve MRAR to some extent, but the discussion of multivariate relationships in the rank swapping literature is typically limited to bivariate correlation coefficients [40] [32]. RWN is fully multivariate, exploiting the theorem in Section 7.

RWN also features several additional forms of randomness. Standard rank swapping perfectly preserves univariate distributions, which is good for data utility. However, this can lead to privacy issues, especially for rare values. For example, consider Example C in Section 2. In this context, a value that only occurs once in the data may be disclosive of the presence of an individual known to match that value, especially if the intruder knows that standard rank swapping was used to protect the published data. If the rare values are extremely large or small values, this could be prevented through top or bottom coding, but other cases may be more subtle.

By contrast, RWN duplicates data values, rather than swapping them. Consequently, some values may appear more or less frequently in the released data than in the original data. It is also possible for some values present in the original data to disappear entirely in the released dataset. RWN also allows for the possibility that some values may be the original ones if the q tuning parameter is set to a value less than 1. These additional forms of randomness create more uncertainty for an intruder.

Standard rank swapping is limited to ordinal and numeric data. As mentioned, some work has been done to develop rank swapping variants for other types of data such as nominal data and partially ordered sets [37] [44]. However, a variant of rank swapping that is able to handle mixed datasets consisting of numeric, ordinal, and nominal data that is widely available does not currently exist. RWN also address this issue.

4.3 Comparison to Microaggregation

Multivariate microaggregation is another thoroughly studied SDC method, available for instance in the **sdcMicro** package via the functions **microaggregation()** and **microaggrGower()**.

The method involves forming small clusters of similar observations and using an aggregation function, typically the mean, to compute a single shared value for each variable in the cluster. Perturbed records are formed by replacing each record in a cluster by the vector of aggregated values corresponding to that cluster.

Numerous variations of the method exist using different approaches for forming clusters and different aggregation functions [34]. Microaggregation has been tested in numerous empirical papers and found to perform favorably in comparison with other SDC methods [12] [11]. Versions of microaggregation that can handle categorical data exist [43] [40].

Though microaggregation clusters are in some sense similar to RWN neighborhoods, there are very substantial differences between the two. Unlike noise addition and most data swapping methods

[2] *Semantic multivariate rank swapping*, a recent variant of rank swapping for nominal data, also takes the approach of defining swapping intervals for each record instead of for each value [37].

including RWN, microaggregation causes the level of variation in the data to be reduced. Also, as will be discussed below in Section 9.1, microaggregation is directly effective in a k-anonymity sense, which will not be the case for these other methods, including RWN.

There are also other differences between RWN and microaggregation. First, the clusters used in microaggregation correspond to groups of records, and the groups do not overlap. For RWN, each record has its own neighborhood, and neigborhoods for different records can overlap with each other. Second, microaggregation allows values that do not exist in the original data to be present in the perturbed data. RWN only duplicates values, so the only values that can appear in the perturbed data are those present in the original data.

Unlike RWN and rank swapping, microaggregation methods are typically deterministic.[3] As discussed above, the randomness in the RWN algorithm gives the intruder more uncertainty when trying to deduce information about individuals in the original data. On the other hand, as pointed out in [12], a fully or partially deterministic approach may be preferable to one in which a different modification is made for each query (ruled out in our context in Section 5), as otherwise an intruder may be able to gain insight about the original data through averaging.

5 Statistical Views and Goals

With any methodology for data analysis, one must first give a clear statement of the context and goals, the topic of this section.

5.1 Released Microdata

We assume the traditional SDC context:

- We have data on which users will perform statistical analysis, yet has a privacy requirement. A typical example would be data on patients, to be used for disease research but which must preserve privacy of the patients.
- In pursuing the privacy requirement, the data curator modifies the data in some manner, then releases the modified data to the users. The modification is made just once. This is in contrast to the typical implementation of differentially private methods; see below.
- Users of the released microdata may then conduct various statistical analyses of whatever type they wish. No distributional assumptions, nor restrictions on statistical methodology, may be assumed by the data curator in implementing a privacy mechanism.

5.2 The Role of Differential Privacy (DP) in Microdata

The preceding subsection, involving the release of modified microdata for use in statistical analysis, is traditionally the typical SDC context. But how, if at all, does DP fit into this context? DP is often referred to as the "gold standard" of data privacy [48] [24] [7], applicable to any dataset, so it is imperative that this question be addressed here.

In the form typically discussed, DP methods are not applicable to the freeform microdata analysis defined above. Instead, many DP methods are defined specifically for a given type of statistical output.

One common introductory example is that of the mean of one of the columns in the data; Laplace-distributed noise is added to the actual mean in the raw microdata, calibrated to the desired DP

[3] A variation of the method where randomness is added is [40].

protection level, and the noisy mean is then presented to the user. A more advanced example is that of the Report-Noisy-Max algorithm, which estimates the index of the maximum value within a set of numbers in a differentially private manner [10]; here noise is added to the individual data points. A related class of methods is *local* DP [16], an extension of *randomized response* survey techniques. Here noise is added at the source of data creation, say a survey respondent.

This means that in DP, a separate DP-compliant method must be developed for each statistical procedure used—mean, regression analysis, visualization tools, etc. This key difference was described by statistician Larry Wasserman [46] (who refers to DP and classic SDC as the "computer science" and "statistics" views),

> CS view: Receive a query for a [specific statistical procedure], return a private answer.
>
> Statistics view: Give me data. Then I can: draw plots, fit models, test fit, estimate parameters, make predictions ...

In other words, Wasserman wants a "free pass" to do a variety of statistical and data-analytic operations. Let's call this Free Pass SDC.

However, in many cases, microdata is amenable to a DP approach. A significant example is that of entirely categorical data, i.e. a contingency table. The multinomial cell counts are then sufficient statistics for any quantity of interest to Free Pass users. Thus a DP-compliant method that, say, adds noise to the cell counts would fit into our context here. Say there are q cells and N_i denotes the result of adding noise to cell i, in a DP-compliant manner. Then by the Composition Theorem [16], the vector $N = (N_1, ..., N_q)$ is DP-compliant. This is in essence the Census Bureau approach. See also [45].

We note too that conditional quantities, say mean wage of Hispanic high school teachers, are then also DP-compliant, because any function of N will have that property, as seen by considering inverse images in the mathematical definition of DP. However, the basic DP criterion may not even be defined in some conditional situations, as will be discussed in Section 6.2.

6 Desirable Characteristics

In developing a new SDC procedure, such as our proposed RWN, these goals were key for us:

- Ability to handle mixed continuous and discrete/categorical data.
- Ability to accept the full range of queries.
- Preservation, to the degree possible, of multivariate distributions/relations, for proper statistical analyses (MRAR property).
- Easy interpretability of the privacy afforded by the method, as perceived by those who are affected.

Let's elaborate on these goals, in the light of RWN and existing microdata-based methods, including differential privacy (DP).

6.1 Handling Mixed Continuous and Categorical Data

A major obstacle to noise addition methods is their inability to handle discrete/categorical data. Consider a variable such as Number of Children in Family. After noise addition, a value may become

negative, an unacceptable situation. Truncation at zero induces bias. A similar difficulty arises with categorical variables, after they are converted to dummy (*one-hot*) form.

Data swapping methods generally cope well here, and RWN in particular handles mixed continuous and discrete/categorical data in a simple, natural, computationally straightforward manner.

In the US Census case, data curators had an additional constraint, in that they wished to preserve certain marginal totals in the contingency table. This was not an issue in terms of the privacy guarantees themselves, as the DP criterion allows *post-processing* [16], though matching all the desired marginal totals necessitated complex and highly time-consuming linear programming algorithms.

6.2 Handling the Full Range of Queries

Cell suppression by its very nature fails this criterion, as seen in the example of the Athens, GA householder in Section 2 above. Surprisingly, DP also suffers from this drawback, again due to its very nature:

The formal definition of DP involves comparing the probabilistic behavior of a query function f on two (real or potential) databases differing only in records x and x'. Say for example that the record x is that of our male Athens householder, and the paired record x' differs in terms of gender. Suppose the only householder with 10 children in Athens is male.

Then since f in this example is a conditional mean, it is undefined on the database containing x', i.e. the value of a query regarding the mean income of all female householders with 10 children is undefined. The definition of DP itself is the problem here, and a differentially private response is impossible. In other words, even a DP-compliant database must resort in part to using non-DP methods, such as cell suppression (which the US Census application of DP does). Note again, as discussed in Section 2, this type of setting has been deemed key throughout the years of research in SDC methods, including the DP literature; it is not a minor anomaly.

As a member of the class of data swapping methods, RWN does not have this problem.

6.3 Preservation of Multivariate Relations

Absent some compensating feature, any change to the data arising from applying a Free Pass SDC procedure (and for that matter, most DP methods) will result in distortions of the relations between variables in the data, typically in the form of attenuation. Since analysis of multivariate relations comprise the very core of statistics, we take as a major goal at least approximately preserving such relations.

We are of course willing to allow the preservation of those relations be just one aspect of the utility/privacy tradeoff that is necessary to any disclosure avoidance technique. Again, we call the utility aspect of this in the context of preserving multivariate relations Multivariate Relations Attenuation Resistance (MRAR). The goal then is to develop an SDC method that includes MRAR, with the method providing the database curator "levers" that allow them to choose their desired utility/privacy tradeoff level.

MRAR is a challenging condition to meet, and thus comparatively little work in the SDC field has focused on it. For instance, it is mentioned only briefly in [25] and [14].

In the noise-addition realm, there has been [26] [28] [41]. If the added noise preserves covariance matrix structure up to a multiplicative constant p, then statistical methods such as linear regression and principal components analysis can be made valid. However, even then, some statistical quantities will be distorted.

In the realm of data swapping methods, pairwise correlations can also be preserved to a user-defined extent with the **rankSwap()** and **shuffle()** functions in the **sdcMicro** [40] package. The correlation level is a tuning parameter.

The theorem presented in Section 7 shows that RWN does offer MRAR.

6.4 User Interpretability of the Privacy Parameters

Consider any database involving human subjects (HSs). (The same comments will apply to corporate subjects and so on.) In choosing a privacy mechanism, it is vital that the database curator win the confidence of HSs that the curator has made all reasonable efforts to preserve the HSs' privacy. Any privacy mechanism will have associated with it various tuning parameters, and it is vital that those HSs understand the nature of the numerical values of those parameters, in the privacy/utility tradeoff.

Data swapping methods, including our RWN method here, excel in this regard. For instance, RWN's parameter q is the probability that a given field in a given record will be swapped. An HS can take comfort in this, thinking (correctly),

> Each of my attributes will be replaced by that of another HS, with probability q. My record in the database will likely not be identifiable as mine, and even if identified, there is probability q that any given value in the perturbed record isn't mine.

In other words, in data swapping methods, the tuning parameters have a simple, direct interpretation. Classical noise-addition methods are also directly interpretable. Going further, an HS in a dataset modified by microaggregation with cluster size k can take solace in knowing there are $k-1$ other HSs who are recorded as exactly like them. (This however, may be mitigated by the number and nature of the variables not involved in the clustering process.) This HS can directly understand how much privacy is afforded them by the parameter k.

The situation is more complicated in the case of DP. The basic notion is clear enough: Under the usual intuitive description of DP, the HS will think,

> The results of database queries will not be much affected by my presence in the database, so I am effectively invisible.

However, the central question is, *how much* protection the HS is provided; just how much is meant by "not much affected"? This is much less clear, as the DP tuning parameter ϵ often has no easy interpretation like say q above. For instance, in the 2020 US Census, $\epsilon = 19.61$ [20]. That figure is quite devoid of meaning for most HSs, who would not have any inkling whether that is "a lot or a little" amount of protection. And any HSs who understand the technical aspects would likely say, "Since $e^{19.61}$ is over 300 million, it's not clear that I have any protection at all."

Worse, recall that small tabular cells have been a recurring theme in SDC research, often the focus of attention. Yet, very large values of ϵ such as 19.61 tend to give reasonable protection only to large cells.

One of the inventors of DP, Frank McSherry, considers any ϵ value of more than 1.0 as problematic, and characterized the value of 14, surmised for Apple DP, as "pointless" [23].

6.5 Ability to Address the Three Canonical Examples in Section 2

A. Intruder has knowledge of certain small cells:

As noted, cell suppression methods are highly problematic here, and DP may be forced to use them as well. The other methods should do well.

B. Intruder knows the identity of some extreme record:

Rank swapping and similar methods will not solve this problem. RWN and microaggregation will handle it well. DP in the multinomial setting described in Section 5.2 may or may not solve the problem.

One common approach by the Census Bureau over the years, *top coding*, would help; here a maximum value is set, say on income, with any value exceeded it to be replaced by the top code.

C. Intruder simply wishes to know whether a given entity (person, firm etc.) is represented in the dataset:

In general, methods having high degrees of k-anonymity such as microaggregation will do well here, almost automatically. Data swapping methods, including RWN, will do well, providing the tuning parameters are chosen to have sufficiently high modification so as to prevent a high probability of record re-identification. This issue also involves the small-cell problem.

6.6 Summary

As noted, data swapping methods generally have all of our listed desirable properties, except for MRAR. As a data swapping method, RWN "inherits" those desirable properties, as well as solving the MRAR problem.

7 Rationale and Theoretical Basis for RWN

Since we are using neighborhoods, one might ask, "Why not just replace entire data rows—a given row is replaced by a neighboring row—rather than do replacement component by component, taking different components from different neighboring rows?" That would achieve our MRAR goal, but at the possible expense of too much increase in the standard errors of estimated quantities of interest to the analyst, and possibly make record re-identification easier for the intruder.

We must ask whether our method has the all-important MRAR property. The following theorem, cast in an idealized setting, shows that it does. Informally:

> Let f_X be a density function for the p-variate vector $X = (X_1, ..., X_p)$. Consider the conditional distribution of X, given that X is in a small neighborhood of a point t. Then the X_i are approximately independent in this distribution.

For expositional convenience, the theorem and proof will be stated for the case $p = 2$.

Theorem: Consider a bivariate random vector (X, Y) having a joint density, and set $\epsilon > 0$. For any t in $\mathcal{R}^2$, let $A_{t,\epsilon}$ denote the ϵ neighborhood of t. Let F denote the joint cdf of (X, Y). Given $(X, Y) = t$, define $G_{t,\epsilon}$ to be the cdf of (X, Y), given that that vector is in $A_{t,\epsilon}$. Finally, given $(X, Y) = t$, define independent random variables U_ϵ and V_ϵ to be drawn randomly from the first- and second-coordinate marginal distributions of $G_{t,\epsilon}$, respectively. Then

$$\lim_{\epsilon \to 0} P\left(U_\epsilon \leq a \text{ and } V_\epsilon \leq b\right) = F(a, b) \tag{3}$$

for all $-\infty < a, b < \infty$.

In other words, as ϵ goes to 0, the bivariate distribution of (U_ϵ, V_ϵ) goes to that of (X, Y), *even though U_ϵ and V_ϵ are independent while X and Y are not independent.*

□

Note that (3) concerns the unconditional distribution of (U_ϵ, V_ϵ). The latter is a random vector in $A_{(X,Y),\epsilon}$.

Intuitively, although U_ϵ and V_ϵ are conditionally independent, the vector (U_ϵ, V_ϵ) is close to (X, Y) in the unconditional distribution and thus has approximately the same joint distribution.

Proof:
First,

$$\lim_{\epsilon \to 0} U_\epsilon = X \tag{4}$$

and

$$\lim_{\epsilon \to 0} V_\epsilon = Y \tag{5}$$

Using the Bounded Convergence Theorem, we have

$$\begin{aligned}
\lim_{\epsilon \to 0} P\left(U_\epsilon \leq a \text{ and } V_\epsilon \leq b\right) &= \lim_{\epsilon \to 0} E\left[P\left(U_\epsilon \leq a \text{ and } V_\epsilon \leq b \mid X, Y\right)\right] && (6)\\
&= \lim_{\epsilon \to 0} E\left[P(U_\epsilon \leq a \mid X, Y) \cdot P(V_\epsilon \leq b \mid X, Y)\right] && (7)\\
&= E\left[1_{X \leq a} \cdot 1_{Y \leq b}\right] && (8)\\
&= E\left[1_{X \leq a \text{ and } Y \leq b}\right] && (9)\\
&= P(X \leq a \text{ and } Y \leq b) && (10)\\
&= F(a, b) && (11)
\end{aligned}$$

8 Neighborhoods and Tuning Parameters

The neighborhoods are formed using both the Euclidean distance-based radius ϵ and the number of nearest neighbors k, which must be specified by the user. In short, ϵ provides control over the similarity of the data points within a neighborhood, while the nearest-neighbor parameter k guarantees that the neighborhood will have sufficiently many data points.

For many datasets, there are typical records as well as records with more unusual or extreme values. Typical records will have many neighbors even for small values of ϵ while unusual records may have zero neighbors unless ϵ is large. The following discussion will use for illustration the **bodyfat** data from the R package **mfp** [3].

If ϵ alone were used to form neighborhoods, RWN would exclude records with empty neighborhoods from the released dataset. This would protect their privacy, which is important considering that these unusual records may correspond to more identifiable individuals, but would result in a complete loss in their utility. To avoid this, the user could increase ϵ until these more extreme records have non-empty neighborhoods. However, as seen in Figure 1, an increase in ϵ can substantially increase the size of the neighborhood for typical records as well, causing the values in a typical record to be mixed with very dissimilar records in the perturbation process, potentially leading to a decrease in utility of the perturbed data.

On the other hand, using k alone would impose both an upper and lower bound on the neighborhood size. For instance, for small k, the neighborhood size may be suitable for unusual records but undesirably small for typical records. Furthermore, a uniform neighborhood size is more easily exploited by an intruder. Thus, having two complimentary neighborhood size parameters ϵ and

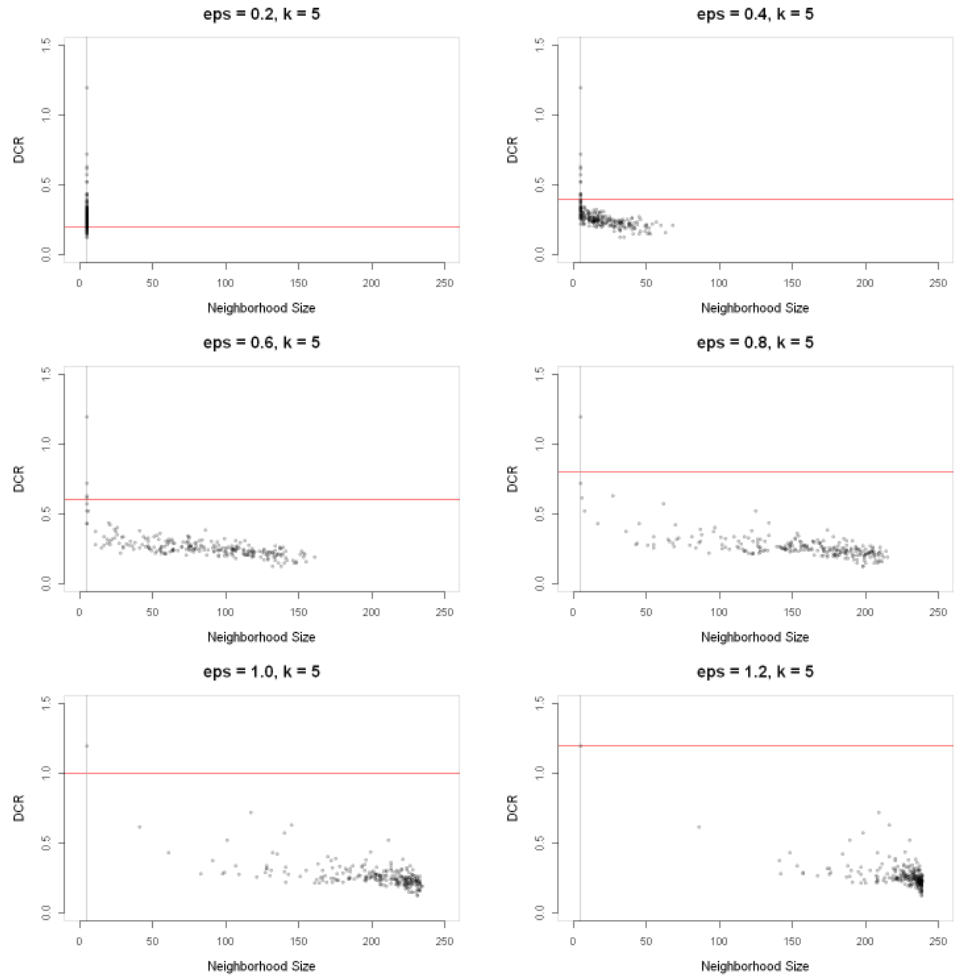

Figure 1: Body Fat Data, Plots of Minimum Distance vs. Neighborhood Size

k gives the database curator finer control over the perturbation of the data to balance utility with privacy for a specific dataset.

In Figure 1 we illustrate how different choices of RWN tuning parameters affect neighborhood size. We calculate the distance to the closest record (DCR) for each record in the original dataset. Here, the distances are calculated between records within the original dataset. Then, we plot distance to closest record against the neighborhood size (i.e. the number of records within the neighborhood) for multiple choices of ϵ while holding constant $q = 1$ and $k = 5$. In the charts, the red horizontal line depicts the value of ϵ and the gray vertical line depicts the value of k. When ϵ is small, the points are clustered along the vertical gray line, reflecting that the k parameter is primarily dictating the size of the neighborhoods. As ϵ increases, many points stray away from the vertical gray line as neighborhood size increases for many records. However, a few points stay near the gray vertical line. These are outlying data points who are ensured to have at least five neighbors since $k = 5$ was chosen.

9 Empirical Investigations: Criteria

Any SDC method is a balance of statistical utility and degree of privacy. Empirical investigation of the method must then define measures for these two criteria. We note at the outset, though, that different SDC methods may require different measures.

9.1 Privacy

Measurement of privacy can be complex. In this section, we explore that issue and relate it to RWN and other methods.

9.1.1 Which Variables Are Modified?

It is common in the SDC literature to distinguish between different kinds of variables such as identifiers, quasi-identifiers, and confidential attributes (see e.g. [38] [35]). Identifers are variables that can uniquely identify individuals on their own such as social security number. These variables must be removed from the released data. Quasi-identifier variables such as ZIP code contain information that generally does not uniquely identify an individual and may be available to an intruder. Individuals may, however, be uniquely identified using combinations of values in multiple quasi-identifers variables. Confidential attributes contain information that intruders do not have access to and may want to learn.

As noted by [38], "any attribute is potentially a quasi-identifier, depending on the external information available to the intruder." In our modern day where vast amounts of data are collected and sold about individuals, distinguishing between quasi-identifiers and confidential attributes is increasingly challenging and unrealistic, especially since the data curator often cannot anticipate what types of information will be possessed by intruders [39].

Consequently, we do not distinguish between quasi-identifiers and confidential variables anywhere in our RWN analysis, and we have RWN perturb all variables in the input dataset. But with many other SDC methods, modification may typically be made only to some variables. This may make it difficult to compare the various methods.

9.1.2 k-Anonymity

One of the most well-known privacy measures, *k-anonymity*, requires that for each unique combination of values in the quasi-identifier variables, there must be at least k records in the dataset sharing that combination of values.

Over the years, various shortcomings and limitations of k-anonymity have been identified, leading to many enhancements such as *p-sensitive k-anonymity*, *l-diversity*, and *t-closeness*. However, these enhancements are also not entirely satisfactory. For instance, data utility is often a major issue for datasets achieving some form of k-anonymity [13].

Microaggregation with cluster size k essentially makes k-anonymity automatic. This is also the case with certain *probabilistic k-anonymity* methods that are special cases of microaggregation [38]. In general, though, k-anonymity will not be a good criterion for assessing methods such as RWN and rank swapping.

On the contrary, these methods will typically result in little or no increase in what we will call *mean k-anonymity* (MKA). We define that as follows. For each record in the dataset, determine the maximal value of k such that this record matches $k-1$ others. MKA is then the average of this value over all records.

Privacy Method	m	MKA
none	5	480.92
rank swap, R0=0.10	5	451.79
rank swap, R0=0.90	5	467.55
RWN	5	483.22
none	25	1699.84
rank swap, R0=0.10	25	1566.30
rank swap, R0=0.90	25	1603.99
RWN	25	1700.49

Table 1: MKA Experiment

Here is a brief numerical look. We consider **svcensus**, a dataset included in the R **qeML** package. This is data on programmers and engineers, from the 2000 US census. MKA was computed as the average value in the **fk** component in the return value of **sdcMicro::freqCalc()**.

There are 20090 rows and 6 columns, the latter for age (numeric), education (ordinal), occupation (6 categories), wage income (numeric), weeks worked (numeric), and gender. All ordinal variables were converted to ordinal numeric. We omitted the income and occupation variables. We ran **rankSwap()** from **sdcMicro**, with $R0 = 0.10, 0.90$, using all columns as key variables. RWN was run with the number of nearest neighbors k set to 10.

Also, we discretized age, first rounding to the nearest $m = 5$ years, then to the nearest $m = 25$. Larger values of m should lead to larger MKA.

The results are presented in Table 1. Neither rank swapping nor RWN achieved a substantial increase in MKA; if anything, MKA seems to degrade under rank swapping. Results for other tuning parameter values, not shown here, were similar. In other words, k-anonymity seems to be a poor choice of privacy measure in some settings.

9.1.3 Record Re-identification

Another class of re-identification risk measures are record linkage-related. This typically takes the form of the intruder determining the closest record in the perturbed data to the intruder's external knowledge of the intended target. Criteria of this form have been widely used in empirical studies comparing SDC methods [12] [35] [34]. One empirical study explains the following:

> The number (or the proportion) of correct re-identications is a common record linkage-based measure of disclosure risk. However, this measure has some limitations that we next discuss. It is certainly appropriate when SDC is achieved by masking the quasi-identier attributes, whereas the sensitive attributes are left unmodified (or are only slightly modified). However, if the sensitive attributes have been signicantly altered, a correct linkage may not be equivalent to disclosure [11].

In our analyses here, we consider two criteria along these lines.

9.1.4 Probability of Self-Closest Record (PSCR)

Consider an "ideal" scenario in which the intruder knows *all* of some individual's attributes. The intruder then finds the perturbed record closest to this attribute vector, and assumes that this is actually the record belonging to the target. We may determine the probability that the intruder successfully re-identifies the target's perturbed record in this manner. This is the setting of Example C in Section 2, but we can still take it as a general privacy criterion.

9.1.5 Distance to Closest Record (DCR)

Another measure in this class is Distance to Closest Record. The computation of DCR involves comparing the original and released datasets by computing the distances between records across the two datasets. Different variations of the measure have been used in the literature. For example, [11] computed DCR using the ranks instead of the data values themselves. In general, a larger DCR value is taken to imply less disclosure risk for the record in question. However, it is important to note that DCR provides an incomplete picture of disclosure risk. For instance, a DCR value of 0 may not imply high disclosure risk if the data contains categorical variables with few unique values or the range of the numeric variables is small. In these cases, the identical record in the released data may be provided some privacy by being "hidden in a crowd" of similar records [29].

9.1.6 Other Measures

Another category of privacy measures is interval disclosure risk. The basic idea is that even if an intruder cannot exactly determine the attribute of an individual, they may still be able to determine a narrow range of values or set of categories that the true value falls in. For example, little privacy is provided when an intruder can determine that a patient's diagnosis is one of several similar conditions or can infer the salary of an employee within a few thousand dollars [16].

One way to compute interval disclosure risk for numeric data is to create an interval around each released data value and then check whether the corresponding value in the original data falls within the interval. A higher proportion of values in the original data within the constructed intervals implies higher disclosure risk. The intervals can be created in various ways such as using standard deviations, ranks, and robust Mahalanobis distance [40] [27].

As mentioned earlier, detecting outliers and rare values is another critical part of any privacy analysis since these individuals may be the most easily identifiable. There are many statistical methods for detecting outliers. For example, outliers may be identified using Mahalanobis distance or via Cook's distance for linear regression analysis [47].[4]

9.2 Statistical Utility

As noted, statistical analysis is at its core a matter of identifying relations between variables. The question to consider in the SDC context is whether the relationships that exist in the original data tend to remain intact in the released data. Here we follow [17], who note that a reasonable measure is to assess whether the released dataset "can obtain approximately the same substantive [relational] results while simultaneously protecting the privacy." For instance, consider the color correlation plots in Figure 5.[5] We find that we can perturb data while retaining broad correlation structure, including to a large extent the strength of the correlations.

Another aspect of utility is validity of standard errors for statistical inference purposes. For SDC methods affecting only a small portion of the data, this is less of an issue. For RWN, one can prove that the standard errors are asymptotically valid, as $q \to 0$ at the proper rate relative to n. (See [17] for a DP solution, under certain assumptions. As usual, the problem of discrete/categorical variables remains a challenge.)

Since we emphasize MRAR, we utilize numerous measures that capture how well multivariate relationships are preserved in the released data. We address this by investigating how correlations, principal components, regression coefficients, and R-squared values are affected by perturbation. We also investigate using those relationships for prediction of new cases.

[4] The measure is based on the effect of leaving-one-out operations, and is thus not a "distance" *per se*.

[5] These use the **bodyfat** dataset (Section 9.3.2).

9.3 RWN Experiments

We now apply these considerations to experiments involving RWN.

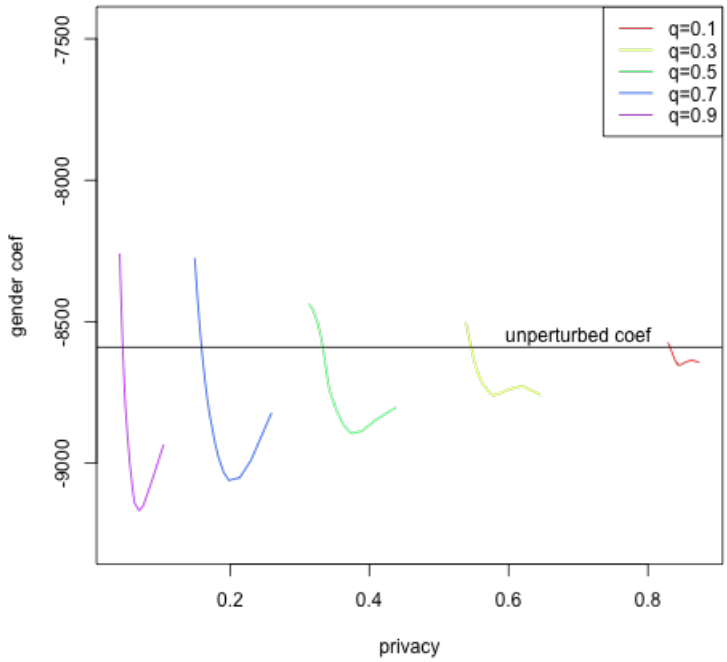

Figure 2: Utility vs. Privacy, for various k

9.3.1 Privacy-Utility Tradeoff

Of course, larger values of q, and of k or ϵ, will lead to poorer statistical utility, the classical Privacy-Utility Tradeoff. Here we consider an example, again based on the **svcensus** dataset.[6] Our measure were as follows:

- Our privacy criterion was Probability of Self-Closest Record (PSCR), defined above in Section 9.1.4.
- For the statistical utility criterion, we took an actual statistical problem: We ran a linear regression analysis, predicting wage income from the other variables. One of the variables is for gender, and we are interested in the estimated coefficient for the indicator variable for being female, which was -$8595.6. Statistical utility then is defined as the closeness of the perturbed coefficient estimate to this figure.

We present several graphs here. In each, the number of nearest neighbors $k = 5, 10, 15, ..., 75$, and q varies as shown in the graph legends. As RWN has a randomness component, the graphs

[6] We use the full dataset here, not deleting some columns as before.

show means over 10 replications; *loess* smoothing was used to produce the curves. Note that (a) the smaller the privacy value, the better, and (b) the closer the utility value is to -8595.6, the better.

- Here we plot utility versus privacy. Each value of k corresponds to a point on the tradeoff curves, though k does not appear explicitly in the graph. The results are shown in Figure 2.
- Next, we plot privacy against k, shown in Figure 3.

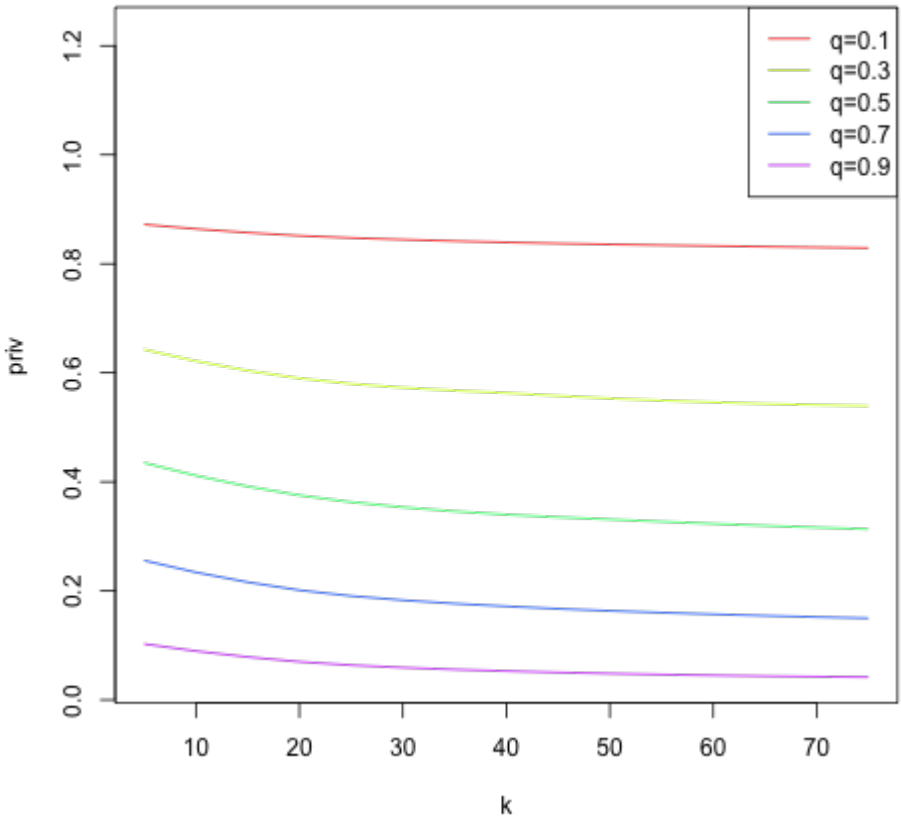

Figure 3: Privacy vs. k

- Finally, we plot utility against k; see Figure 4.

The graphs show the standard trend in SDC settings: The lesser privacy levels yield the greater statistical accuracy.

Figure 3 is somewhat surprising. Though, as expected, greater values of k yield better privacy, the dominant factor is q rather than k.

Figure 4 is possibly the most interesting one. Though we see quite substantial effects of q, they rather converge at k around 50 or 60.

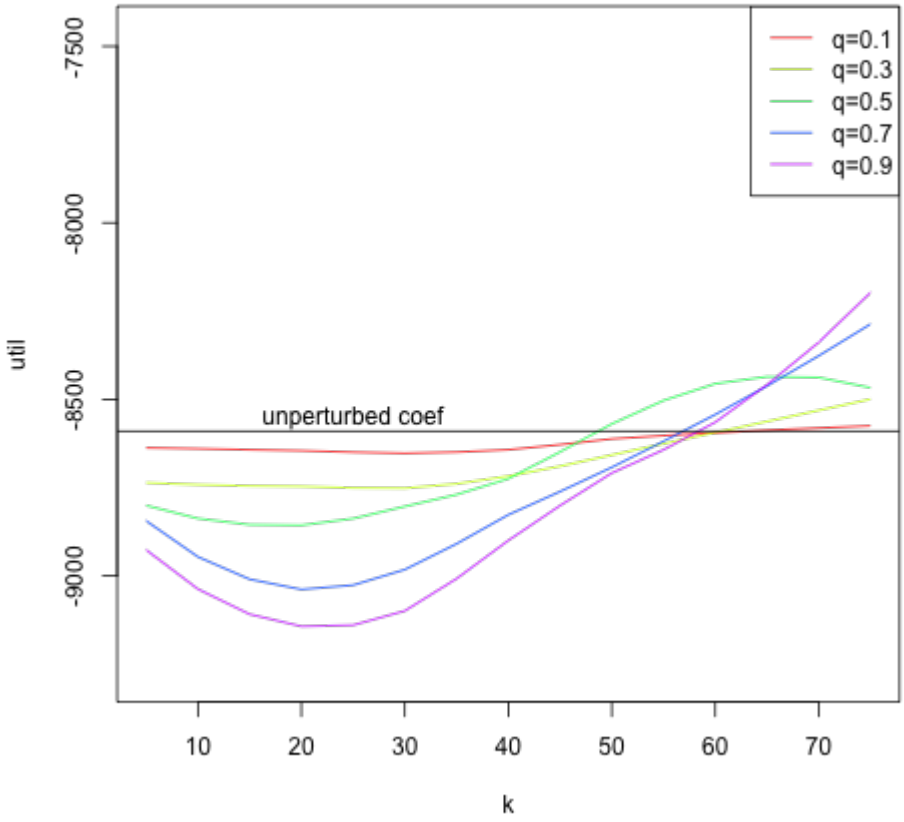

Figure 4: Utility vs. k

9.3.2 Bodyfat Data

The **bodyfat** dataset [4] consists of measurements of 252 adult males. Two estimates of their body fat percentages are calculated using the Brozek and Siri equations. For our analysis, we use only the body fat percentages based on Siri, eliminate 11 observations that contain values that appear to be erroneous or biologically implausible such as body fat percentages less than four percent, and calculate Body Mass Index (BMI) for each individual.

For the experiments in this section, we use the cleaned **bodyfat** dataset as well as several perturbed versions of this dataset created using RWN with tuning parameters $k = 5$, $q = 1$, and varying values of ϵ from 0 to 1.2 by increments of 0.05.[7] We conduct a variety of analyses to illustrate the performance of RWN through a variety of visualizations and scenarios. For these analyses, we generate only one perturbed dataset using RWN for each ϵ value. These experiments build intution about proper tuning parameter selection as some combinations of tuning parameters achieve a more optimal trade-off of privacy and utility than others. In each figure where a horizontal blue line is

[7] The neighborhood of a point is determined by k or ϵ, whichever produces a larger neighborhood.

present, the line denotes the result for the unperturbed data.

Correlation:

We calculated Pearson correlation coefficients pairwise for all variables in the **bodyfat** dataset. The colored correlation matrices in Figure 5 show that for small ϵ, the correlation coefficients in the perturbed and unperturbed data for a given variable are of the same sign and of very similar magnitude. This provides evidence of the MRAR property of RWN. As ϵ increases, the correlation coefficients tend towards zero. Thus, for large ϵ, the variables in the perturbed data are essentially uncorrelated.

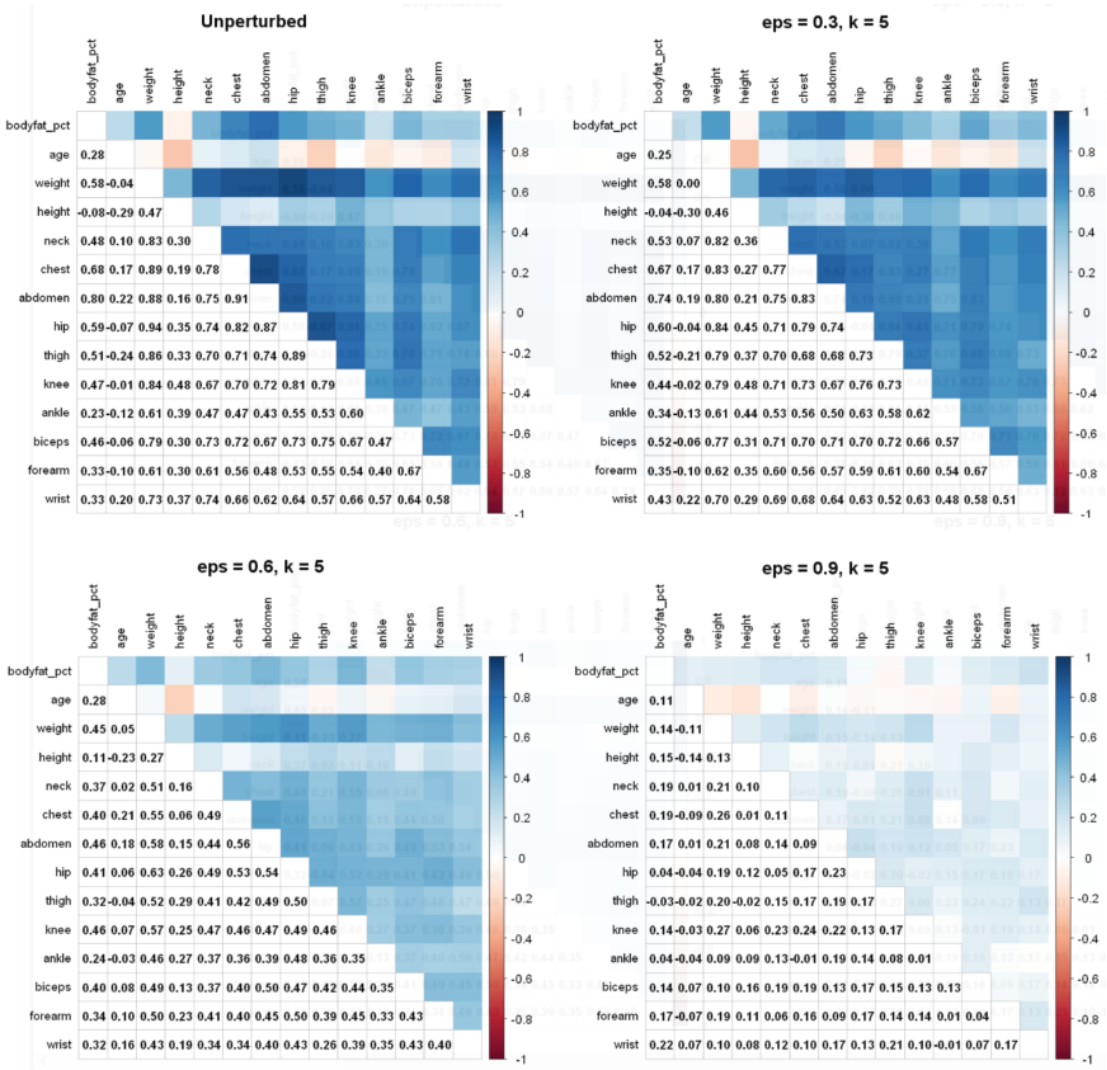

Figure 5: Body Fat Data, Correlation Matrices for RWN

Principal component analysis:

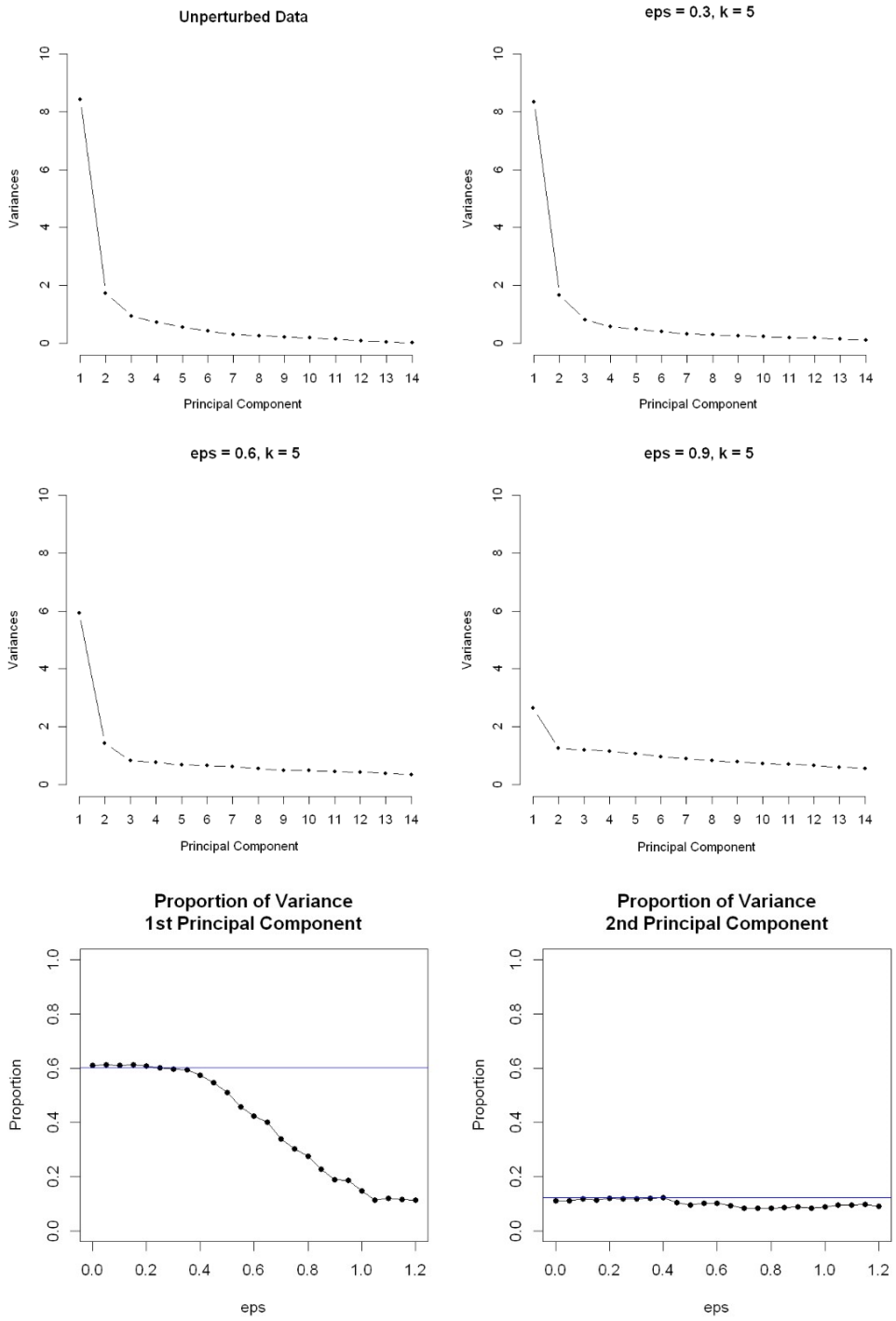

Figure 6: Body Fat Data, PCA Scree Plots and Proportion of Variance Plots

Obs. Number	Body Fat Pct.	BMI	Age	Weight	Height	Neck	Chest	Abdomen	Hip
37	35.2	48.9	46	363.2	72.3	51.2	136.2	148.1	147.7
39	34.5	39.1	45	262.8	68.8	43.2	128.3	126.2	125.6
205	47.5	37.6	51	219.0	64.0	41.2	119.8	122.1	112.8
231	35.0	33.9	65	224.5	68.3	38.8	119.6	118.0	114.3
168	29.9	33.2	37	241.3	71.5	42.1	119.2	110.3	113.9

Table 2: 5 Individuals in Body Fat Data with Highest BMI

After scaling the data, we performed principal component analysis (PCA) on both the unperturbed and perturbed datasets. In Figure 6, we see that the majority of the variance in the data is along the first principal component. For small values of ϵ, the scree plot is very similar to the original data, but as ϵ increases, the variation is spread over more principal components. This is consistent with Figure 5 where we observed that most of the body measurements are positively correlated, but the correlations go to zero as ϵ increases.

Outliers:

We regressed body fat percentage on all other variables including BMI, and computed Cook's distances. One use of Cook's distance is to identify areas of the space spanned by the explanatory variables where few observations are found.

The cleaned bodyfat dataset contains one individual who is substantially larger than all the others. Consequently, this individual would be one of the most easily identifiable in the unperturbed data. Table 2 shows the five individuals with the highest BMI values and a subset of their body measurements. The person who stands out in size is Observation 37.

In Figure 7, this individual corresponds to the largest Cook's distance in the unperturbed data. However, even with minimal perturbation, the maximum Cook's distance becomes much lower, suggesting that this individual is no longer as easily identifiable in the perturbed data.

We can also visualize how perturbation affects this individual through plots of principal component scores. This outlier is highlighted in each plot in Figure 8 using a red triangle while other data points are represented using black diamonds. In the unperturbed data, the outlier has the largest score on the first principal component by far. In the perturbed data, the outlier retains the largest score on the first principal component but stands out much less noticeably from rest of the data.

Returning briefly to Figure 1, we see there is one record which has the same neighborhood of size 5 for all values of ϵ. This record corresponds to the exceptionally large individual. Consequently, although we plot maximum Cook's distance and principal component scores for a variety of ϵ values, the perturbed records corresponding to this individual in the various perturbed datasets are just different randomizations based on the same neighborhood determined by $k = 5$. However, the other data points are affected by the increasing ϵ, which is seen in the shrinking size of the cloud of points due to the first principal component accounting for a lesser proportion of the total variance.

9.4 Empirical Comparison of SDC Methods

Next, we use the **bodyfat** data to compare the empirical performance of RWN with rank swapping and multivariate MDAV microaggregation as implemented by **sdcMicro** through the **rankSwap()** and **microaggregation()** functions. Since RWN and rank swapping both involve randomness and may generate different datasets each time they are run, we generate 100 distinct datasets for each choice of tuning parameters to ensure that the comparative performance of these methods is not dependent on a single anomalous run of these algorithms. For these methods, we compute each performance measure separately using each of the 100 released datasets and then take the average to

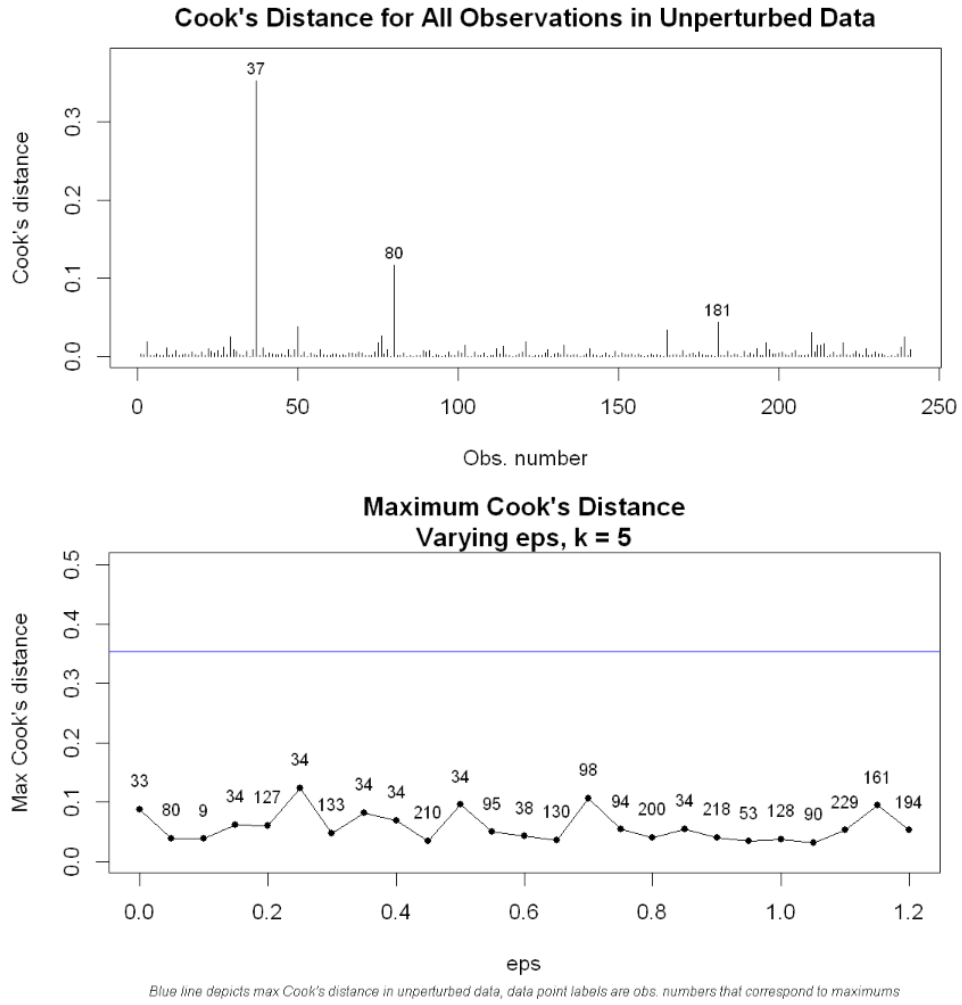

Figure 7: Body Fat Data, Cook's Distances

obtain a single value for a particular choice of tuning parameters. The version of microaggregation used is deterministic, outputing the same released dataset when given the same input data and cluster size, so the microaggregation results are based on a single dataset for each cluster size.

A fundamental issue for comparative analyses of different SDC methods is finding a proper basis of comparison. What choices of tuning parameters are comparable? Examining the structure of the algorithms themselves does not provide a definitive answer. For example, RWN's neighborhoods are analogous to the swapping ranges in standard rank swapping, but RWN's neigbhorhoods are allowed to be of widely varying sizes while swapping ranges are of uniform size, so there is no exact equivalence between the two. Microaggregation defines disjoint clusters of records while RWN defines non-disjoint neighborhoods for each record in the dataset. Consequently, there is no clear equivalence between neighborhood size and cluster size.

Many empirical analyses in the literature compute a *score*, a weighted average of various disclosure risk and information loss measure, to compare the empirical performance of each method across a

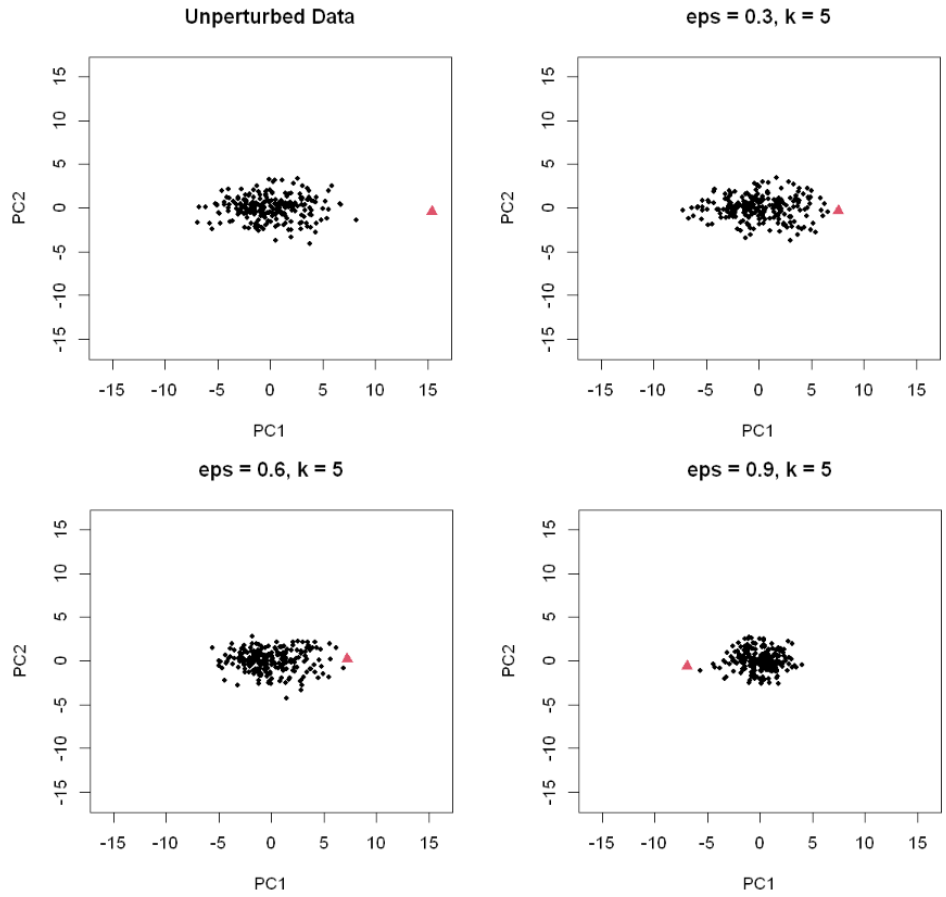

Figure 8: Body Fat Data, PCA Outlier

variety of tuning parameters selections [12] [35] [34]. Our analysis is inspired by [11] which uses a variation of DCR to find tuning parameters for each method that provide a comparable level of privacy when applied to a specific dataset. Then, using the chosen tuning parameters, each method is compared using a utility measure to determine the best performing method. As seen their analysis, the relative performance of each method may differ when applied to different datasets. We follow this general procedure with numerous modifications.

First, we identify numerous choices of tuning parameters that are comparable in terms of DCR. We use a variation of DCR where distances are calculated using the values in the data (instead of using ranks) and the minimum Euclidean distance is found for each record in the original dataset. Let v denote the number of replications. $v = 100$ for RWN and rank swapping, and $v = 1$ for microaggregation. Let W denote the original dataset with records $r_1, \ldots, r_n$ and let $W^{'(k)}$ denote the perturbed datasets with records $r_1^{'(k)}, \ldots, r_n^{'(k)}$.

$$DCR(r_i) = \min_j ||r_i - r_j^{'(k)}||_2$$

Tuning Parameter Index	RWN (ϵ)	Rank Swapping (P)	Microaggregation (k)
1	0.25	0.12	2
2	0.30	0.15	3
3	0.35	0.18	4
4	0.40	0.21	5
5	0.45	0.24	6
6	0.50	0.27	7
7	0.55	0.30	8
8	0.60	0.33	9
9	0.65	0.36	10
10	0.70	0.39	11
11	0.75	0.42	12
12	0.80	0.45	13
13	0.85	0.48	14
14	0.90	0.51	15
15	0.95	0.54	16

Table 3: Tuning Parameter Values for Comparison

$$Risk(W, W^{'(k)}) = \frac{1}{nv} \sum_{k=1}^{v} \sum_{i=1}^{n} DCR(r_i)$$

As in the previous section, for RWN we use $k = 5$ and $q = 1$, so the only RWN tuning parameter that varies is ϵ. The tuning parameter index column in Table 3 is used as the x-axis in all the comparison plots in this section.

In Figure 9, it is immediately clear that all three methods are comparable for indices 3 through 7. However, other comparable values can be found by drawing horizontal lines through the chart. For instance, similar DCR values are attained for RWN with $\epsilon = 0.80$ (index 12), rank swapping with $P = 0.54$ (index 15), and microggregation with $k = 14$ (index 13).

We also use the interval disclosure measure from **sdcMicro** to examine these disclosure risk results from a different angle. For standard deviation interval disclosure risk, we use a standard deviation of 1. In Figure 10, we see that using standard deviation-based intervals, RWN achieves a much lower level of disclosure risk. For robust Mahalanobis distance-based intervals, the difference is much less distinct but the relative performance remains the same. These results suggest that the privacy equivalencies established using DCR are conservative for RWN. That is, they may understate the level of privacy provided by RWN relative to other methods.

Next, we compare the performance of the three SDC methods in terms of information loss measures. The first information loss measure IL_1 is the mean absolute difference of Pearson correlation coefficients. Let X and $X^{'(k)}$ denote the correlation matrices corresponding to the original and perturbed datasets W and $W^{'(k)}$, respectively.

$$IL_1(X, X^{'(k)}) = \frac{2}{vp(p-1)} \sum_{k=1}^{v} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} |x_{i,j} - x_{i,j}^{'(k)}|$$

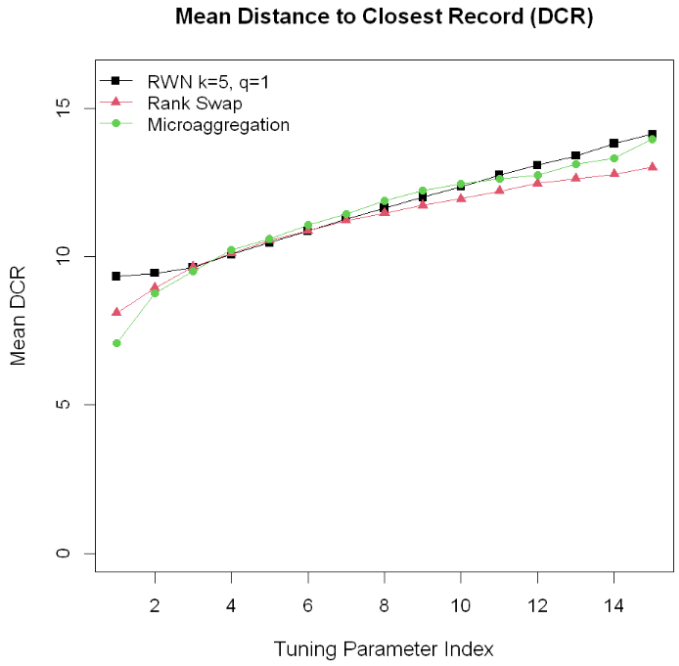

Figure 9: Body Fat Data, Distance to Closest Record Comparison

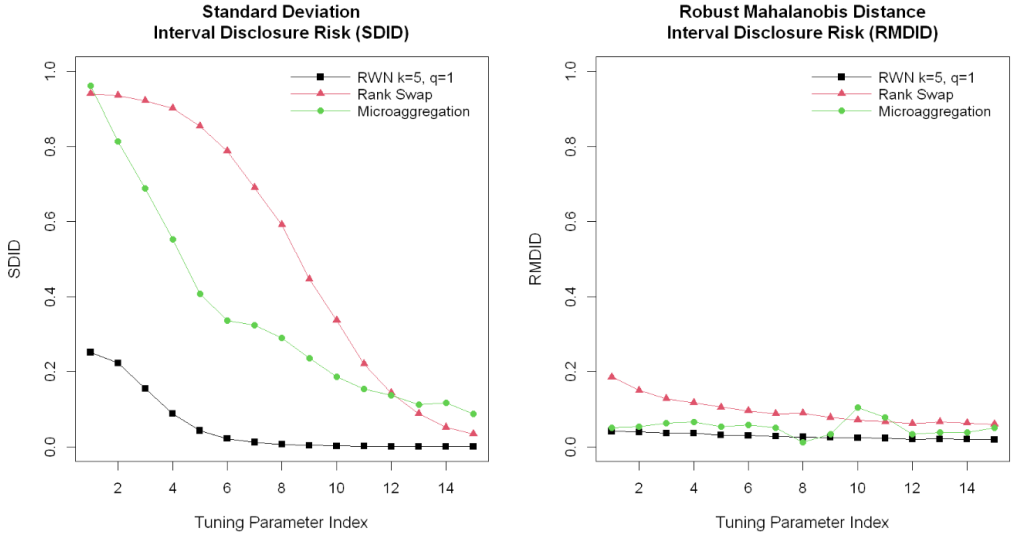

Figure 10: Body Fat Data, Interval Risk Comparison

For this measure, a higher value implies more information loss. In Figure 11, we see that RWN and rank swapping slightly outperform microaggregation at lower DCR values, while the latter is substantially better for higher DCR values.

Note, though, that the absolute differences between correlation coefficients do not provide a full

picture. The algebraic sign of the difference is also important. Microaggregation tends to strengthen correlations as the level of perturbation increases while rank swapping and RWN tend to weaken correlations as the level of perturbation increases. This phenomenon can be seen by comparing Figure 12 with Figure 5. Ideally, correlations should neither increase nor decrease upon perturbation. In MRAR terms, we see that microaggregated data will tend to overestimate relations between variables, while RWN will tend to underestimate them.

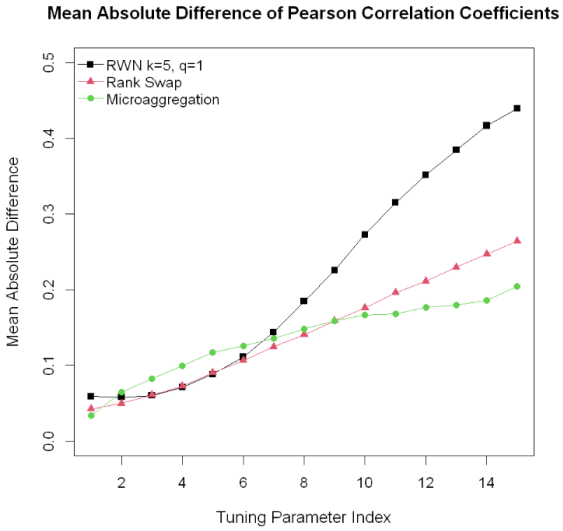

Figure 11: Body Fat Data, Correlation Comparison

The second utility measure IL_2 is the mean absolute difference of PCA loadings. Considering the scree plots in Figure 6, we only analyze the loadings on the first and second principal components. We analyze the loadings one principal component at a time. Let Y and $Y^{'(k)}$ denote the PCA loadings matrices corresponding to the original and perturbed datasets W and $W^{'(k)}$, respectively. Let $Y_{.j}$ and $Y_{.j}^{'(k)}$ denote the loadings for the j^{th} principal component.

$$IL_2(Y_{.j}, Y_{.j}^{'(k)}) = \frac{1}{vp} \sum_{k=1}^{v} \sum_{i=1}^{p} |y_{i,j} - y_{i,j}^{'(k)}|$$

In Figure 13, we see that all three methods perform very comparably for the first principal component, but for higher DCR values, RWN performs substantially less favorably for the second principal component.

The third utility measures IL_3 is the mean absolute difference of regression coefficients. Let β and $\beta^{'(k)}$ denote the vectors of estimated regression coefficients corresponding to the original and perturbed datasets W and $W^{'(k)}$, respectively.

$$IL_3(\beta, \beta^{'(k)}) = \frac{1}{vp} \sum_{k=1}^{v} \sum_{i=1}^{p} |\beta_i - \beta_i^{'(k)}|$$

In Figure 14, we see that rank swapping and RWN attain similar levels of information loss for IL_3. The line for microaggregation is much more jagged, likely due to multicollinearity and the fact that $v = 1$. Unlike RWN and rank swapping, microaggregation strengthens the correlations as the cluster

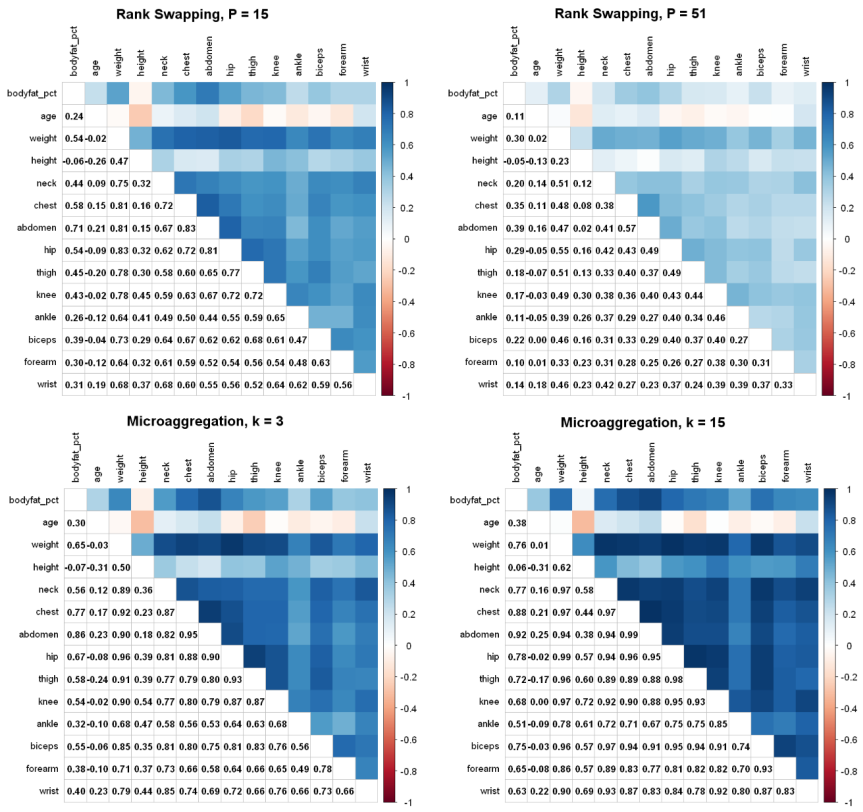

Figure 12: Body Fat Data, Correlation Matrices for Rank Swapping and Microaggregation

size increases. Consequently, increased cluster size leads to more severe multicollinearity, and the jaggedness in the microaggregation line reflects this.

We also examine the R-squared values. The R-squared values for RWN and rank swapping follow a similar pattern: as neighborhoods and swapping ranges increase in size, R-squared values decrease. This is consistent with our observation of correlation coefficients in Figures 5 and 12. In contrast, the R-squared values for microaggregation increase as the cluster size increases because variation in the data is lost to a greater degree for larger cluster sizes. As in the correlation comparison, RWN and rank swapping perform very similarly for smaller tuning parameter choices, but rank swapping achieves less information loss than RWN for larger values of ϵ.

In summary, we find that RWN performs most similarly to rank swapping when applied to the **bodyfat** data. For this dataset, RWN outperforms rank swapping in terms of interval disclosure risk but performs similarly or slightly worse in terms information loss in multivariate relationships depending on the tuning parameter selection. It should be noted that the relative performance of these methods may differ when applied to different datasets, and a data curator could use a similar battery of tests to select an SDC method and find an optimal choice of tuning parameters for their context.

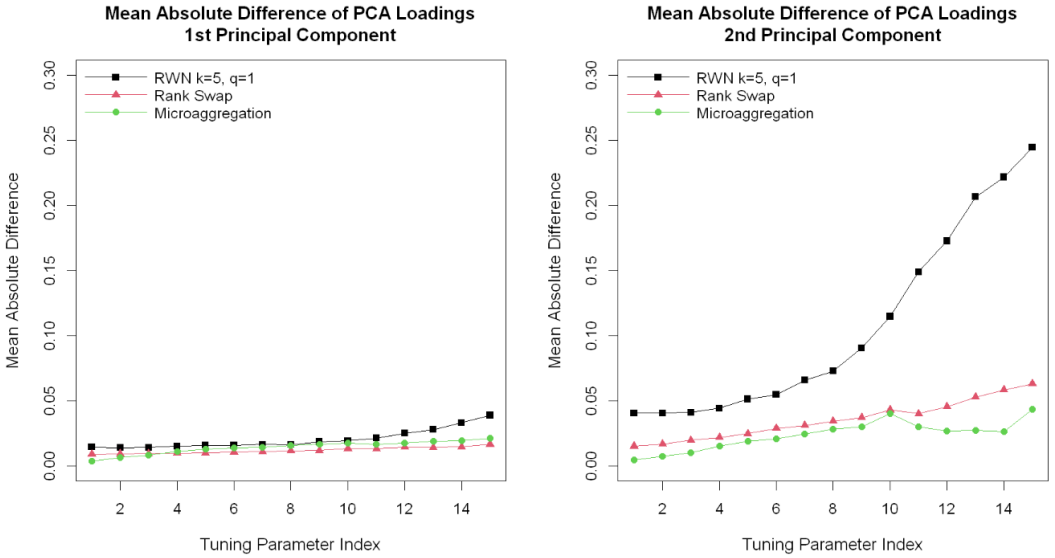

Figure 13: Body Fat Data, PCA Loadings Comparison

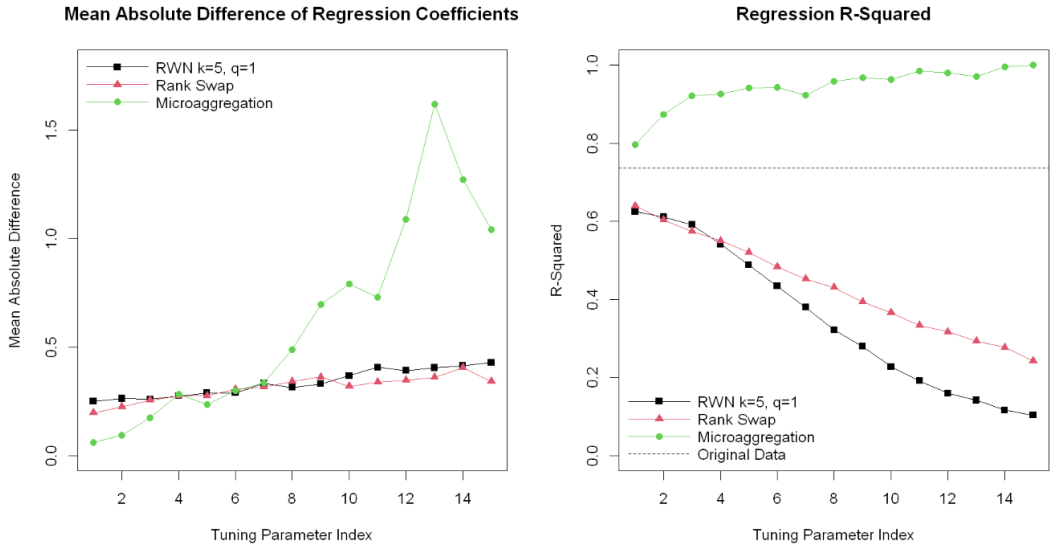

Figure 14: Body Fat Data, Regression Comparison

9.5 Prediction-Oriented Assessment

Statistical applications tend to fall into one of two general categories, which we will refer to as Description and Prediction. The former has the goal of understanding some entity, process, effect and so on, while the latter concerns predicting new data.

Much of the SDC literature has been aimed at the Description side of things, estimating means, totals, regression coefficients and the like. The Prediction side has rarely been the focus, and we now turn to that aspect in this section. We are interested particularly in classification settings, the

form of many modern applications.

Intuitively, SDC methods should do fairly well in classification settings. Consider 2-class problems, for instance, where we are predicting $Y = 0, 1$ from a vector of covariates X, say with a continuous distribution. Assuming equal misclassification costs, the set of t for which $P(Y = 1|X = t) = 0.5$ forms the decision boundary. What is the effect of RWN's data perturbation?

Consider a nonparametric regression method, say random forests. If a data point in the training set is far from the boundary, perturbation will have little or no effect on future predictions; only points very close to the boundary would be likely to experience change in prediction. One may conjecture, then, that SDC methods will not compromise prediction ability much, at least for nonparametric regression methods. This was confirmed in our experiments.

We first consider **pef**, a dataset included in the R **regtools** package. This is data on programmers and engineers, from the 2000 US census. We predict a variable **occ**, which codes one of six occupations, from variables such as age, income and education. Here we took $q = 0.5, \epsilon = 0$, and varied $k = 5, 10, 25, 50$. Misclassification rates, using random forests, are as follows:

Random Forest Misclassification Rate				
Unperturbed	k = 5	k = 10	k = 25	k = 50
0.628	0.627	0.645	0.682	0.659

This dataset does not lend itself to strong predictability, with an error rate about about 63%. However, that rate increases only slightly under RWN. The above results were based on 25 replications, with a holdout set size of 1000.

Here is the same analysis on a second dataset, the well-known Pima diabetes study. It's quite different from the census data, in that it is much smaller (768 rows, vs. 20090 for **pef**), thus requiring more privacy protection. On the other hand, greater predictive accuracy is possible for this data. Again, performance appears not to decline due to the privacy action, and may even help, due to salutary smooting effects.

Random Forest Misclassification Rate				
Unperturbed	k = 5	k = 10	k = 25	k = 50
0.241	0.242	0.233	0.245	0.241

We briefly investigate this phenomenon using the **pef** dataset. Total correlation (also known as multiinformation) is an information theoretic measure of the multivariate relationships in the data [30]. Since the **pef** dataset contains a mix of numeric and categorical variables, many of the exploratory experiments used for the **bodyfat** data like Pearson correlation and principal component analysis cannot be applied to the entirety of the **pef** dataset. Total correlation is the Kullback-Leibler divergence between the joint distribution $p(X_1, X_2, \ldots, X_n)$ and the product of the marginal distributions $p(X_1)p(X_2)\ldots p(X_n)$. A larger total correlation implies a greater degree of dependence among variables whereas a total correlation near 0 implies that the variables are nearly independent.

After discretizing numeric variables, we calculate total correlation using one instance of the **pef** dataset perturbed by RWN for each of the tuning parameter combinations listed above. The results are displayed in the table below.

Total Correlation				
Unperturbed	k = 5	k = 10	k = 25	k = 50
1.133	1.181	1.176	1.146	1.133

Interestingly, the degree of dependency among variables in minimally perturbed datasets is slightly higher than in the original dataset. As expected, this degree of dependency decreases as the level of perturbation increases.

10 Large Data Sets and Computational Complexity

10.1 Current Trends

Data sets are growing rapidly in both size and dimension. This is driven by many factors, including:

- A proliferation of data sources, including applications, back-office systems like ERP, smart devices, and sensors.
- The growth of the systems themselves, as businesses and other organizations achieve global scale.
- Increases in data collection and storage capacity through networks and mass storage.
- Larger computing devices and cloud scale computing.
- The transition from recording data about entities (such as people) to storing transactions like purchases to storing events like clicks. Rather than understanding people, the goal now is often how people behave in terms of transactions, mobility and so on.
- The transition from structured to semi-structured to unstructured data. Structured data, like that found in traditional database tables, is strictly constrained in terms of dimension. Semi-structured data like JSON objects, which is often captured from running applications, is unconstrained in dimension. Unstructured data, like documents, photos, or videos, is of almost unlimited dimension. Unstructured data might not seem like a candidate for the method described in this paper, but documents, photos, and videos are often reduced to a set of features described by categorical, binary, or numeric variables.

As a result, datasets can now easily be in the billions of rows with hundreds or thousands of variables or features. But with the growth of datasets comes a growth in risk. Larger datasets put more people at risk, and increased dimensionality increases damage per person. Also, the increase in dimensionality makes it easier to identify someone's record in a dataset. We need to be able to apply the method to these huge datasets, so the algorithm must be reasonably efficient from a computational complexity perspective. As we will see in the next section, the basic algorithm is computationally expensive, but with slight modification can handle large datasets.

10.2 Computational Complexity of RWN

To assess computational complexity of the method, going forward we will assume p is large but fixed, $q = 1$, and n is increasing. We will also assume that the distance measure used to define closeness is arbitrary. This is reasonable, since we need wide latitude to define closeness in different ways for very different datasets. Under these conditions, whether we are selecting nearby points based on epsilon neighborhood or k-nearest neighbor, the computational complexity of the basic method is $O(n^2)$, since we have to calculate the distance between all possible points. This is probably not tenable for large datasets, e.g., $n = 10^9$, for which the number of calculations would be of the order 10^{18}.

10.3 Alternative Methods

10.3.1 Method 1: Draw Neighbors from a Sample of Points

For the first method, we will just take a smaller sample of data points to use as neighbors. That is, we will take a sample S of size $m << n$. For each data point in the original dataset, select the neighbors from the sample, then apply the method as before. For this method, the complexity is $O(mn)$, considerably better than the original method. For $n = 10^9$, $m = 10^4$, complexity is of order 10^{13} instead of 10^{18}, an improvement of five orders of magnitude. Assuming that all the data is in memory, the cost of generating a sample of size m is $O(m)$, so if we were to use a distinct random sample for each point, the complexity would still be $O(mn)$. For $n = 10^9$, $m = 10^4$, complexity is still order 10^{13}. The advantage is that we potentially get slightly richer data and better protection.

10.3.2 Method 2: Sample Randomly from the Distance Matrix

In Method 1, we are actually sampling twice as many points as we need to, since when we sample point x_j to be a possible neighbor of the point x_i, we can also use x_i as a potential neighbor of x_j. This is due to the symmetry of distance measure and the distance matrix $D = [d_{ij}]$ where d_{ij} is the distance between points i and j. So instead of drawing a new sample for each data point, we could simply draw a sample of elements from the entire distance matrix. Then for each data point, we could look at the distances in the sample and find those that are sufficiently small or find the smallest k elements.

We will now describe the method more formally and derive the complexity. Let $W = (x_{ij})$ be the original dataset. We want to sample pairs (x_i, x_j), $i < j$ at random with a sample size of $n_s = nm/2$ where $m << n$ is the desired sample size per data point. There are $n(n-1)/2$ such pairs total, and we are sampling $n_s << n(n-1)/2$. To generate the sample, we can use a mixed congruential random number generator to generate numbers in $[1, n(n-1)/2]$.

Once we have n_s random integers in the range $[1, n(n-1)/2]$, we will map them to (i, j) pairs, $1 <= i < j <= n$ to obtain the set $S_d = \{(i, j)\}$. We then calculate the distances $d_{ij} = < xi, xj >$ and store them in an undirected graph G_d (for efficient retrieval) with nodes representing the n data points and edges $\{e_{ij} : (i, j) \in S_d\}$. Since the complexities of generating the random numbers, mapping the random numbers to (i, j) pairs, calculating distances, and finding neighbors are all $O(nm)$, the complexity of the entire algorithm is still $O(nm)$.

10.3.3 Method 3: Partitioning

With the advent of cloud computing, it has become much more feasible to harness the power of many computers or virtual machines, each with gigabytes of memory. Under this scenario, the problem (and data) are partitioned and spread across multiple instances, and then the results are combined to form a single modified dataset. There are several ways this approach could be applied to the problem at hand:

1. Simply partition the dataset into u equal partitions and then apply the original method. In this case, the size of each partition is n/u, so the complexity is $O(un^2/u^2) = O(n^2)/u$, a reduction in overall complexity by a factor of u over the original method. However, if the partitions are processed in parallel, the complexity for each individual partition represents a reduction of a factor of u^2 over the original method.

2. Partition the dataset into u equal partitions and then apply Method 1 or 2 above. Again, the size of each partition is n/u but we are only calculating distances for a sample of size m, so the overall complexity is $O(umn/u) = O(nm)$, the same as it is for Method 1 or 2 without

partitioning. However, if the partitions are processed in parallel, the complexity for each individual partition is $O(mn/u)$, a reduction by a factor of un/m over the original method.

Partitioning has two potential advantages: It spreads the work across multiple instances, thereby increasing the computing power that can be brought to bear on the problem, and in the case of approach 1 above, it further reduces the complexity of the problem being solved on each node. However, with partitioning, the resulting combined dataset will be somewhat different.

11 Conclusion

We have presented a new SDC method for data release, distinguished by its ability to preserve multivariate relations as well as handle mixed continuous and discrete/categorical data, and provided a theoretical basis for the role of neighborhoods in achieving MRAR.

In addition to its theoretical appeal, RWN possesses practical features needed by database curators. RWN's tuning parameters k, ϵ, and q provide curators multiple levers to adapt the method to a specific dataset and context. The curator must select and tune an SDC method according to the special characteristics of the database at hand: Prevalence of outliers, distributional traits of the variables, presence of small cells, a need to make some parts of the data more secure than others, and so on. In deciding what kind of privacy must be provided, the curator may need to take into account requirements of public policy specific to their domain or their organization's legal contracts. Hence, the importance of a particular measure of utility or privacy may differ by context. In light of this, we performed a variety of statistical analyses to examine privacy and utility from multiple angles. For example, a curator who aspires to provide special privacy to individuals who are easily identifiable due to having unique or rare attributes may choose to adopt our Cook's distance approach. Our empirical results demonstrate RWN's ability to balance the twin goals of preserving data utility while providing privacy.

References

[1] ABOWD, J., ASHMEAD, R., CUMINGS-MENON, R., GARFINKEL, S., HEINECK, M., HEISS, C., JOHNS, R., KIFER, D., LECLERC, P., MACHANAVAJJHALA, A., MORAN, B., SEXTON, W., SPENCE, M., AND ZHURAVLEV, P. The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*, Special Issue 2 (June 2022). https://hdsr.mitpress.mit.edu/pub/7evz361i.

[2] ABOWD, J. M., AND SCHMUTTE, I. M. Economic Analysis and Statistical Disclosure Limitation. *Brookings Papers on Economic Activity 46*, 1 (Spring) (2015), 221–293.

[3] AMBLER, G. *mfp: Multivariable Fractional Polynomials*, 2015. R package version 1.5.2.

[4] AMBLER, G., AND BENNER, A. *mfp: Multivariable Fractional Polynomials*, 2023. R package version 1.5.4.

[5] BRAND, R. *Microdata Protection through Noise Addition*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 97–116.

[6] CARLSON, M., AND SALABASIS, M. A data-swapping technique for generating synthetic samples : A method for disclosure control.

[7] CHAUDHURI, K., IMOLA, J., AND MACHANAVAJJHALA, A. *Capacity Bounded Differential Privacy*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[8] DENNING, D. E., AND SCHLORER, J. Inference controls for statistical databases. *Computer 16*, 07 (1983), 69–82.

[9] DePersio, M., Lemons, M., Ramanayake, K. A., Tsay, J., and Zayatz, L. n-cycle swapping for the American Community Survey. In *Privacy in Statistical Databases* (Berlin, Heidelberg, 2012), J. Domingo-Ferrer and I. Tinnirello, Eds., Springer Berlin Heidelberg, pp. 143–164.

[10] Ding, Z., Kifer, D., E., S. M. S. N., Steinke, T., Wang, Y., Xiao, Y., and Zhang, D. The permute-and-flip mechanism is identical to report-noisy-max with exponential noise. https://arxiv.org/abs/2105.07260, 2021.

[11] Domingo-Ferrer, J., Ricci, S., and Soria-Coma, J. Empirical comparison of anonymization methods regarding their risk-utility trade-off. United Nations Economic Commission for Europe, 2017.

[12] Domingo-Ferrer, J., and Torra, V. *A Quantitative Comparison of Disclosure Control Methods for Microdata*. North Holland, 2001, p. 111–133.

[13] Domingo-Ferrer, J., and Torra, V. A critique of k-anonymity and some of its enhancements. In *2008 Third International Conference on Availability, Reliability and Security* (2008), IEEE, pp. 990–993.

[14] Duncan, G., Elliot, M., and Salazar, G. *Statistical Confidentiality: Principles and Practice*. Statistics for Social and Behavioral Sciences. Springer New York, 2011.

[15] Dwork, C. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II* (Berlin, Heidelberg, 2006), ICALP'06, Springer-Verlag, p. 1–12.

[16] Dwork, C., and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science 9*, 3-4 (2014), 211–407.

[17] Evans, G., King, G., Schwenzfeier, M., and Thakurta, A. Statistically valid inferences from privacy protected data. https://gking.harvard.edu/dp, 2021.

[18] Fienberg, S. E., and McIntyre, J. Data swapping: Variations on a theme by Dalenius and Reiss. In *Privacy in Statistical Databases* (Berlin, Heidelberg, 2004), J. Domingo-Ferrer and V. Torra, Eds., Springer Berlin Heidelberg, pp. 14–29.

[19] Fischetti, M., and Salazar-González, J.-J. Partial cell suppression: A new methodology for statistical disclosure control. *Statistics and Computing 13*, 1 (Feb. 2003), 13–21.

[20] Garfinkel, S. Comment to Muralidhar and Domingo-Ferrer (2023) – Legacy statistical disclosure limitation techniques were not an option for the 2020 US Census of Population and Housing. *Journal of Official Statistics 39*, 3 (2023), 399–410.

[21] Garfinkel, S., Abowd, J. M., and Martindale, C. Understanding database reconstruction attacks on public data. *Commun. ACM 62*, 3 (feb 2019), 46–53.

[22] Gong, R., Groshen, E. L., and Vadhan, S. Harnessing the Known Unknowns: Differential Privacy and the 2020 Census. *Harvard Data Science Review*, Special Issue 2 (June 2022). https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

[23] Greenberg, A. How one of Apple's key privacy safeguards falls short. *Wired Magazine* (Sep. 2017).

[24] Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., and Roth, A. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium* (2014), pp. 398–410.

[25] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K., and de Wolf, P. *Statistical Disclosure Control*. Wiley Series in Survey Methodology. Wiley, 2012.

[26] Kim, J. J. Method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the Section on Survey Research Methods (JSM)* (1986), pp. 303–308.

[27] Mateo-Sanz, J. M., Sebé, F., and Domingo-Ferrer, J. Outlier protection in continuous microdata masking. In *Privacy in Statistical Databases* (Berlin, Heidelberg, 2004), J. Domingo-Ferrer and V. Torra, Eds., Springer Berlin Heidelberg, pp. 201–215.

[28] Matloff, N. S. Another look at the use of noise addition for database security. In *1986 IEEE Symposium on Security and Privacy* (1986), pp. 173–173.

[29] Mendelevitch, O., and Lesh, M. D. Fidelity and privacy of synthetic medical data.

https://arxiv.org/abs/2101.08658, 2021.

[30] MEYER, P. E. *infotheo: Information-Theoretic Measures*, 2022. R package version 1.2.0.1.

[31] MOORE, R. Controlled data-swappng techniques for masking public use microdata sets. https://www.census.gov/srd/CDAR/rr96-04_Controlled_DataSwapping.pdf, 1996.

[32] MOORE, R. A. Controlled data-swapping techniques for masking public use microdata. *Statistical Research Division Report Series RR96-04* (1996).

[33] MURALIDHAR, K., AND SARATHY, R. Data shuffling: A new masking approach for numerical data. *Management Science 52*, 5 (2006), 658–670.

[34] NIN, J., HERRANZ, J., AND TORRA, V. On the disclosure risk of multivariate microaggregation. *Data & knowledge engineering 67*, 3 (2008), 399–412.

[35] NIN, J., HERRANZ, J., AND TORRA, V. Rethinking rank swapping to decrease disclosure risk. *Data & Knowledge Engineering 64*, 1 (2008), 346–364. Fourth International Conference on Business Process Management (BPM 2006) 8th International Conference on Enterprise Information Systems (ICEIS' 2006).

[36] OBERSKI, D. L., AND KREUTER, F. Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review 2*, 1 (1 2020).

[37] RODRIGUEZ-GARCIA, M., BATET, M., AND SÁNCHEZ, D. Utility-preserving privacy protection of nominal data sets via semantic rank swapping. *Information Fusion 45* (2019), 282–295.

[38] SORIA-COMAS, J., AND DOMINGO-FERRER, J. Probabilistic k-anonymity through microaggregation and data swapping. In *2012 IEEE International Conference on Fuzzy Systems* (2012), pp. 1–8.

[39] SWEENEY, L. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10*, 5 (Oct. 2002), 557–570.

[40] TEMPL, M., KOWARIK, A., AND MEINDL, B. Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software 67*, 4 (2015), 1–36.

[41] TENDICK, P. Optimal noise addition for preserving confidentiality in multivariate data. *Journal of Statistical Planning and Inference 27* (1991), 341–353.

[42] TENDICK, P., AND MATLOFF, N. A modified random perturbation method for database security. *ACM Trans. Database Syst. 19*, 1 (Mar. 1994), 47–63.

[43] TORRA, V. Microaggregation for categorical variables: A median based approach. In *Privacy in Statistical Databases* (Berlin, Heidelberg, 2004), J. Domingo-Ferrer and V. Torra, Eds., Springer Berlin Heidelberg, pp. 162–174.

[44] TORRA, V. Rank swapping for partial orders and continuous variables. In *2009 International Conference on Availability, Reliability and Security* (2009), IEEE, pp. 888–893.

[45] WANG, Z., AND REITER, J. P. Post-processing differentially private counts to satisfy additive constraints. *Trans. Data Priv. 14*, 2 (2021), 65–77.

[46] WASSERMAN, L. A statistical view of differential privacy. www.cs.cmu.edu/afs/cs/usr/wing/www/class/15-895/LarryWasserman.pdf, 2011.

[47] WEISBERG, S. *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley, 2014.

[48] XIAO, H., AND DEVADAS, S. Towards understanding practical randomness beyond noise: Differential privacy and mixup. Cryptology ePrint Archive, Paper 2021/687, 2021.