[1]

EFFECTIVE PRIVACY AFTER ADJUSTING FOR INVARIANTS WITH APPLICATIONS TO THE 2020 CENSUS

ROBERT ASHMEAD, DANIEL KIFER, PHILIP LECLERC, ASHWIN MACHANAVAJJHALA, AND WILLIAM SEXTON

Abstract. Differential privacy is a mathematical tool for protecting the confidentiality of records belonging to individuals. One of the key premises of differential privacy is that *any* measurement based on the confidential data must be altered with carefully chosen random noise before publication to mask the possible effect of one person on the data. However, the deployment of differentially private disclosure limitation technologies by official statistical agencies may not always occur under ideal conditions. For instance, internal decisions or external requirements (e.g., legal or contractual obligations) may stipulate that certain statistics must be published *exactly*. Additionally, overlapping datasets may have already been published. In this paper, we explain (1) the semantics of algorithms that satisfy differential privacy, (2) how the semantics are affected by release of exact statistics (computed directly from the confidential data), (3) how to attribute responsibility for any resulting information leakage, (4) how to provide privacy semantics for the combined information leakage.

1. Introduction

The Decennial Census of Population and Housing produces data that are used for apportioning the House of Representatives, redistricting all Congressional and state legislative bodies, intercensal population estimates, public policy, allocation of federal resources, and general research. Historically, the Decennial Census has produced more than 100 billion of statistics about a population whose size is measured in the hundreds of millions. Impossibility results, such as those of Dinur and Nissim [DN03], imply that these statistics cannot be released too accurately without risking a reconstruction of almost all of the original confidential micro-data used to produce those statistics. To control the possibility and accuracy of such reconstruction attacks, the U.S. Census Bureau plans to deploy a differentially private disclosure avoidance system for the 2020 Census of Population and Housing. That system will allow published tables to be accompanied by formal privacy guarantees.

Differential privacy [DMNS06], a notion of algorithm stability, provides a measure of information leakage and enables techniques for controlling the worst-case information leakage about any record. The bound on information leakage is controlled by a parameter ϵ, known as the *privacy-loss budget*. The information leakage bound is determined by comparing the behavior of the algorithm for any pair of input databases that only differ in a single record. Attractive properties of differentially private algorithms include: (a) the privacy guarantee degrades slowly over multiple releases, as opposed to other statistical disclosure control techniques that can allow reconstruction of parts of the underlying data with as few as two releases [GKS08]; (b) it controls the ability of an attacker to *infer* sensitive properties of records even in the presence of side information [KM12, KM14, KS15, NSW$^+$17]; and (c) these guarantees continue to hold even when the source code of the algorithm is made public. Differentially private algorithms require that any computations whose input is a confidential database must be injected with random noise before they are published or otherwise released. In particular, no deterministic computation that uses the confidential data in a non-trivial manner can be published in a differentially private system.

However, *exact* statistics about the data can easily be leaked through other avenues. For example, other data sets and local knowledge can provide information about the total population or number of housing units and group quarters facilities in many census blocks (such information can include real-estate websites like Zillow and unstructured data like Google StreetView). Furthermore, interpretations of Constitutional and

[1]The views expressed in this technical paper are those of the authors and not those of the U.S. Census Bureau. This draft contains no sensitive data, but it has not undergone full internal Census Bureau peer review. Please address comments to william.n.sexton@census.gov.

Date: July 2019.

statutory requirements sometimes indicate that exact statistics—the unaltered tabulation of some components of the confidential database in its final, edited form—be released. For example, the U.S. Supreme Court (Department of Commerce v. United States House of Representatives, 1999) confirmed that the Census Act (13 U.S.C. Section 195) prohibits the use of "the statistical method known as 'sampling' " for apportionment of the House of Representatives. Taken in concert with the one-person, one-vote rule (U.S. Supreme Court, Reynolds v. Sims, 1964), the redistricting amendments to Title 13 incorporated in PL94-171 (1975), and the amendments preceding the 2000 Census (Pub. L. 105–119, title II, § 209, Nov. 26, 1997, 111 Stat. 2480), "the number of persons enumerated without using statistical methods must be publicly available for all levels of census geography which are being released by the Bureau of the Census."[1] In Utah v. Evans (2002) the U.S. Supreme Court clarified that the prohibited "statistical method" in the Census Act is "sampling" and not other statistical methods such as imputation or disclosure limitation. Specifically, statistical methods other than sampling that change the number of persons in each state for the purposes of apportionment, are not prohibited by 13 U.S.C. Section 195. Such methods, for example count imputation and unduplication, have been used for decades.

The implications of these statutes and judicial rulings for statistical disclosure limitation are not clear (e.g., which population subtotals must remain unperturbed). Historical practice at the Census Bureau has been to use disclosure limitation methods that do not change population totals at any geography [U.S02]. The theory underlying these methods originated in the 1970s [Fel72] when the formal privacy analysis of the implications of publishing exact statistics had not yet been discovered. The Census Bureau is now considering various interpretations of allowable statistical methods for use in the 2020 Census disclosure limitation system and their implications on which population statistics must remain unperturbed.

Exact statistics computed from the confidential or sensitive micro-data are called *invariants*. For the 2018 End-to-End Census Test, the invariants initially under consideration included block-level[2] total population, block-level voting-age population counts, the number of group quarters facilities in each block, the number housing units in each block, the number of occupied housing units in each block, and some characteristics of the group quarters facilities (see [Lec19] for the actual End-to-End implementation details). Because of the basic 2020 Census definitions, the number of occupied housing units is equal to the number of householders[3] in the block. The number of vacant housing units is equal to the number of housing units minus the number of occupied housing units. The non-voting-age population is equal to the total population minus the voting-age population. Thus, the block-level householder, vacant housing unit, and non-voting age population would also be invariants by construction.

The choice of invariants for the 2020 Census publications will be set by policy, not engineering. These policies imply that Census Bureau must control all data publications (such as tabulations analogous to those appearing in PL94-171, Summary File 1 and Summary File 2 [Bur12]) in a manner that does not allow reconstructions of the protected data. This requires the Bureau to study the privacy implications of forcing disclosure avoidance systems to respect different sets of invariants; that is, to compare the resulting degradation in privacy guarantees to the ideal setting (without invariants) of a differentially private algorithm with a pre-specified privacy loss budget. Our analysis shows that privacy-loss accounting is materially changed by the presence of invariants. Not only do invariants leak information on their own, they also weaken the semantic guarantees of disclosure limitation methods (including differential privacy) – a phenomenon we call amplification of privacy leakage (also known as composition [GKS08] in the literature). This paper establishes that differential privacy allows this amplified privacy leakage to be controlled by appropriate settings of the privacy-loss budget.

In this paper, we explain the semantics of differential privacy in the ideal setting without invariants. Next, we study the information leakage that results when invariants are added to the differentially private algorithms and quantify upper bounds on the leakage due to the algorithms and the amplification due to the invariants. Finally, we apply these analyses to the invariants that are under consideration for the 2020 Census.

---

[1] `https://www.law.cornell.edu/uscode/pdf/lii_usc_TI_13.pdf`, page 41

[2] The Census block is the smallest unit of geography for which statistics are published. In 2010, there were 11,078,297 blocks with an average population of approximately 30 people per block.

[3] The "householder" is "Person 1" on the Census questionnaire. This is usually the person who supplied the information for all persons living in the housing unit.

While differential privacy has many semantic interpretations, the main interpretation pursued in this document is now known as posterior-to-posterior semantics and is based on ideas that originated in cryptography [KS14, NSW+17, DN10]. Intuitively, it compares an attacker's gain in inference arising from two scenarios—one in which a respondent truthfully reports a record and another in which the respondent reports a fictitious record. We have adopted this approach and properly specialized it to mechanisms that combine invariants with differential privacy.

The introduction of invariants complicates the analysis of privacy. Invariants may not initially cause a privacy breach, but they create an unstable system in which a small amount of extra information could lead to a privacy breach. It can be explained by analogy to curve fitting. Suppose we have a secret 9^{th} degree polynomial f. There are infinitely many such polynomials. If we pick 9 points $x_1, \ldots, x_9$ and release the value of the polynomial at each of them, then there are still infinitely many polynomials that are consistent with $f(x_1), \ldots, f(x_9)$. However, this creates an unstable situation as releasing the value of f evaluated at one more point will determine f completely. Releasing the 10^{th} point by itself would not have been disastrous, but in the context of the other nine points it was. One can say that the release of the first 9 points *amplified* the risk of releasing the 10^{th} point. In a similar way invariants, set up a system of equations that amplify the disclosure risk of releasing additional (even noisy) information. Thus we are interested in answers to the following questions.

(1) If there is a breach that leaks sensitive information, how much of it is due solely to the disclosure avoidance system? In Section 5 we show how obtain an upper bound on this part of the privacy risk.
(2) How can one quantify both the leakage due to the disclosure avoidance system and the amplification effect due to invariants? This style of analysis is discussed in Section 6 and Section 9.
(3) What properties of the disclosure avoidance system can mitigate the amplified risk? The results presented in this paper show that the special type of randomness used in differential privacy is one such property. The amplified risk multiplies the nominal privacy loss budget ϵ by a constant factor that depends on the invariants. Thus the risk can be mitigated by dividing the privacy loss budget of the algorithm by this constant.

Our conclusions are the following:

- The initial privacy leakage due to the disclosure of the invariants is generally difficult to quantify using modern privacy analyses because it heavily depends on what an attacker knows (and there is no consensus about what is reasonable or unreasonable to assume). However, the main concern is how it amplifies the risks of subsequent data releases.
- The upper bound on privacy leakage *solely* attributed to algorithms satisfying differential privacy can be quantified and is the same as if there were no invariants.
- An upper bound for the amplified leakage can be obtained for the invariants under consideration for the 2020 Decennial Census. More precisely, different facts about a record are afforded different levels of protection, whereas without constraints, all facts about a record are protected equally well. This is a consequence of invariants—even simple invariants such as block-level population totals exhibit this behavior. Intuitively, these detailed block population invariants make it harder to protect the reported location of respondents but other information that cannot be inferred from location is still protected.

In Section 2, we describe the components of the Census of Population and Housing data, relying primarily on the schema that is used for the 2018 End-to-End Census Test. In Section 3, we present differential privacy and its properties in the absence of invariants. In Section 4, we take a brief digression to provide a short summary of privacy guidelines that have led to this focus on differential privacy. When invariants are present, in Section 5, we show how to quantify the information leakage that is attributable to differentially private disclosure control algorithms. In Section 6, we develop results for analyzing the amplified privacy leakage of differentially private algorithms that is caused by the invariants. In Section 7, we apply these results to analyze the privacy loss amplification that could be caused by the invariants under consideration for the 2020 Census of Population and Housing. In Section 8, we generalize the posterior-to-posterior semantics of Section 6 to allow more nuanced privacy guarantees for situations which are not covered by the semantics of Section 6. In Section 9, we provide another interpretation of these posterior-to-posterior semantics in

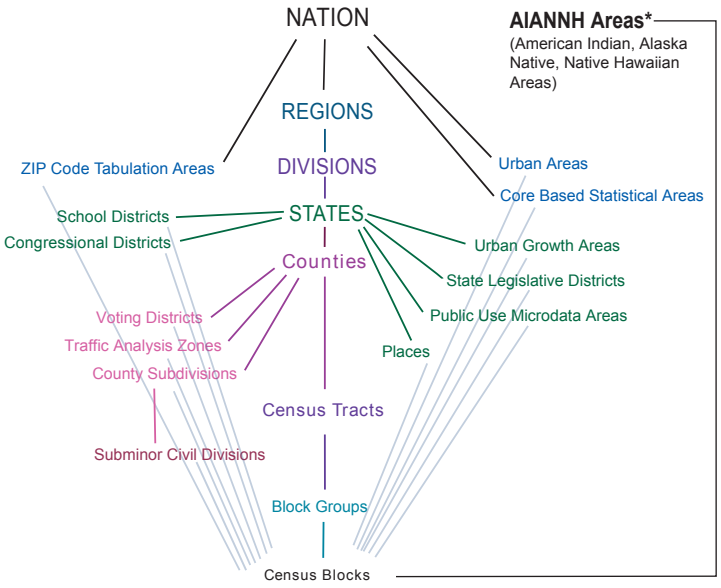

FIGURE 1. Standard Hierarchy of Census Geographic Areas [Bur12]

terms of odds ratios. With these refined semantics, we provide more nuanced guarantees for the proposed invariants in Section 10.

2. Characteristics of the Data

The decennial census collects information about a variety of different entities. These include individuals, housing units, households, group quarters facilities, and geographies.

2.1. **Data Description.**

2.1.1. *Geography.* The fundamental unit of geography is a tabulation block, which is used to create the full geographic hierarchy, including voting districts and other larger areas. Although geography is technically a lattice (see Figure 1), its main structure is a hierarchy:

- The entire United States
- 52 states and state-like entities, which include 50 states, the District of Columbia, and Puerto Rico.
- Counties and equivalent subdivisions (some states do not use the term "county")
- Tracts
- Block groups
- Blocks

Geographic information about blocks, such as whether they fully or partially contain water, are part of urban areas, are part of rural areas, etc, are collected but not considered private. The block boundaries are designed in collaboration with the states and are based on knowledge of the residents in those regions. This complication is ignored and the analysis treats the boundaries as independent of the realized census data. Enumeration geography[4] is pre-specified, and so is fixed prior to the collection of realized data. Tabulation geography[5] is not pre-specified and may depend on results of the census-taking process.

2.1.2. *Group Quarters (GQ).* Group quarters are structures whose primary purpose is to house unrelated people. These include correctional institutions, military barracks, college dormitories, etc. Due to Census Bureau edit rules, a group quarters has its own attributes, such as age restrictions imposed by the edit rules that fix errors in responses to the census questionnaire (for instance, a minimum age restriction for residents of a nursing home would fix an error in which a resident of a nursing home fills in an age of 3),

[4] i.e. the boundaries used for data collection

[5] i.e., the boundaries used for publishing data

	INSTITUTIONAL GROUP QUARTERS
	Correctional Facilities for Adults
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
	Juvenile Facilities
201	Group Homes for Juveniles (Non-Correctional)
202	Residential Treatment Centers (Non-Correctional)
203	Correctional Facilities Intended for Juveniles
	Nursing Facilities/Skilled-Nursing Facilities
301	Nursing Facilities/Skilled-Nursing Facilities
	Other Institutional Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals With Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities With Assigned Patients
405	Residential Schools for People With Disabilities
	NONINSTITUTIONAL GROUP QUARTERS
	College/University Student Housing
501	College/University Student Housing
	Military Quarters
601	Military Quarters
602	Military Ships
	Other Noninstitutional Facilities
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disasters
904	Religious Group Quarters and Domestic Violence Shelters

TABLE 1. Three-digit Group Quarters types [Bur12]

and restrictions on sex (e.g., all-male, all-female, or unrestricted college dormitories). Group quarters have a variety of properties:

- Age restrictions
- Single-sex status (all-female, all-male, no sex restriction on residence)
- High-level type (i.e., institutionalized, non-institutionalized)
- 3 digit type code (see Table 1).

GQ Type 904 is considered extremely sensitive and is treated separately using methods that are confidential. There is an obligation to protect some characteristics of the group quarters. Every group quarters facility must have at least one resident.

2.1.3. *Housing units.* A housing unit is a structure where a family, or other group of related individuals, could (but does not have to) reside. The distinction between a housing unit and a group quarters can be subjective. A housing unit has a vacancy status (vacant or not) and a tenure (e.g., owned, rented, etc.).

Every non-vacant housing unit must have a householder, also known as Person 1, who is the adult resident of the housing unit to whom the other residents are related via their relationship codes.

2.1.4. *Housing status.* There is no official definition for a place where an individual might live. We define housing status as whether (1) a respondent lives in a household and is a householder, or (2) lives in a household and is not a householder, or (3) lives in group quarters.

2.1.5. *Individuals.* Every person resides either in a GQ or a household. Information collected about individuals includes:

- Sex
- Age
- Relation to householder (e.g., householder, spouse of, child of, parent of, step-child of, etc.). For people living in a group quarters facility, their relation to householder attribute is the facility's 3-digit GQ code.
- Hispanic origin (1997 Office of Management and Budget Hispanic ethnicity): a binary attribute (yes or no). In the raw data, more detailed information (such as country of origin) is also recorded.
- Races. There are 6 major race categories, defined in the 1997 OMB standard (White; Black or African American; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; and Other). Individuals can belong to any non-empty subset of them. We treat race as 6 binary attributes with a structural zero (an impossible combination disallowed by edit rules): all race variables cannot be simultaneously 0. In the raw data, more detailed information, such as country of origin, or hand-written race/tribe name is also recorded.

2.2. **Input Data.** The input data consist of the person-level records of every individual, including the code for the block in which their housing unit is located. Block-level records include counts of vacant and non-vacant housing units in each block and the characteristics of group quarters in that block. In the implementation for the 2018 End-to-End Census Test, the relationship of individuals to the householder is not tabulated. The main reason for this is that the End-to-End Test was designed to produce only a subset of the planned publication tables—the PL94-171 redistricting data—and not the full set of summary tabulations. In the End-to-End Test, the full set of structural zeros—e.g. a householder cannot have more than 4 grandparents – is documented in [Lec19].

2.3. **Output Data.** The output ω is a set of records having the same schema as the input data. Thus all of the invariants that are present in the input data D can also be computed from ω. We let Q denote an algorithm that computes the values of the invariants (either from D or ω). We require that the output ω be constrained so that $\mathrm{Q}(D) = \mathrm{Q}(\omega)$.

3. Privacy in the Ideal Setting

Differential privacy [DMNS06] is a cryptographically-inspired privacy definition that is used to design *mechanisms* (i.e., algorithms that protect records in a database). It is designed to hide an individual's contribution to any published statistics. It is also designed to be transparent—the privacy parameters and the source code of differentially private mechanisms can be released without compromising privacy. In contrast, methods like data swapping can become insecure when algorithm source code and privacy parameters (i.e., the swapping rate) become public.

Before reviewing the mathematical definition of differential privacy, we will first discuss privacy semantics in the concrete case of randomized response [War65], a technique for collecting sensitive information in face-to-face surveys. This provides a natural setting for introducing posterior-to-posterior semantics of privacy mechanisms (developed in [KS15, NSW+17]) in the ideal setting where there are no invariants.

3.1. **Randomized response and posterior-to-posterior guarantees.** In the context of surveys, individuals have three basic choices: to participate accurately, to not participate at all, and to withold or falsify some of their information. Incentives for non-participation and falsifying records include the time it takes to answer survey questions truthfully and privacy concerns about answering those questions. The field of survey design can address the first disincentive. Disclosure avoidance technology can address the second. One of the earliest disclosure avoidance solutions was *randomized response* [War65], which predates differential privacy

but is also the first known algorithm that satisfies differential privacy. Under one of the most common variations of this methodology, a respondent answers a yes/no question truthfully with some probability (say 0.51) and otherwise provided a false answer (e.g., with probability 0.49). The odds that the respondent answered truthfully are $0.51/0.49 \approx 1.04$ meaning that the data collector has a large degree of uncertainty about any information of the respondent (odds of 1 equal perfect uncertainty).

Even though the randomized response protocol probabilistically alters a respondent's record, a respondent may submit a fake response anyway. To analyze whether there is any benefit of doing so, consider one respondent—Respondent A—and suppose that Respondent A's truthful answer is "yes". Consider the following two scenarios—one where Respondent A inputs the truthful answer into the randomized response protocol and another where Respondent A inputs some default value into the randomized response protocol. For this example, let us assume that the default value is "No".[6] Let E_t (respectively, E_f) denote the event that Respondent A decides to input the the truthful answer (respectively, default value) into the randomized response protocol. Let O_y (respectively, O_n) be the event that the protocol produces the output "yes" (respectively, "no"). Then, we see that:

$$\frac{P(O_y \mid E_t)}{P(O_y \mid E_f)} = \frac{0.51}{0.49} \approx 1.04$$
$$\frac{P(O_n \mid E_t)}{P(O_n \mid E_f)} = \frac{0.49}{0.51} \approx 0.96$$

In other words, the output of the protocol is likely to be the same, regardless of whether Respondent A uses a real or default record in the randomized response protocol, and hence there is less incentive to falsify data for privacy reasons.

A Bayesian interpretation can be added on top of these semantics. Consider an attacker with a prior belief θ about Respondent A and the general population. This prior could incorporate side information—for instance, the attacker may be a neighbor and so might know the sex and age of A, and might also know statistical information such as how age is correlated with how people answer the survey questions. We impose no restrictions on θ. Let r represent the true attribute of Respondent A. Then simple calculations show:[7]

$$\frac{0.49}{0.51} \leqslant \frac{P_\theta(r = \text{“}yes\text{”} \mid O_y, E_t)}{P_\theta(r = \text{“}yes\text{“} \mid O_y, E_f)} \leqslant \frac{0.51}{0.49}$$
$$\frac{0.49}{0.51} \leqslant \frac{P_\theta(r = \text{“}no\text{”} \mid O_n, E_t)}{P_\theta(r = \text{“}no\text{“} \mid O_n, E_f)} \leqslant \frac{0.51}{0.49}$$

In other words, no matter what the output is (O_y or O_n) and no matter what prior the attacker is using, the posterior inference about Respondent A changes very little even if a default value was used as input to the randomized response protocol. This type of interpretation of privacy is known as posterior-to-posterior semantics.

3.2. **Differential Privacy.** The development of differential privacy has demonstrated two important consequences. First, the privacy guarantee of randomized response can be extended to a wider variety of mechanisms when the data collector is trusted (i.e., when the respondents give their true responses to the data collector and the data collector runs the mechanism on the data). Second, differential privacy enables the use of mechanisms that are much more statistically efficient. For example, randomized response can be used to estimate the true number of "yes" answers among n respondents with standard deviation proportional to $\sqrt{n}$. On the other hand, with differential privacy, this number can be estimated with standard deviation that is constant with respect to n.

There are many variants of differential privacy. The most relevant variant to the 2020 Census is *bounded* ϵ-differential privacy:

[6]This default value could be specified by the survey designer, so it is completely unrelated to the true record of Respondent A.

[7]These calculations are simplified versions of the corresponding calculations for differential privacy, which we show later in this section.

Definition 1 (Bounded DP [DMNS06])**.** A mechanism M satisfies *bounded*[8] *ϵ-differential privacy* if for every pair of databases $D, D' \in X^n$ such that D and D' differ by the modification of one record, and every set of outputs $S \in range(M)$, we have:

$$Pr(M(D) \in S) \leqslant e^{\epsilon} Pr(M(D') \in S)$$

where the probability is only taken with respect to the randomness in M (and not with respect to the data).

The ϵ is known as the *privacy-loss budget*. In the previous randomized response example, the privacy-loss budget was $\log \frac{0.51}{0.49}$ (where log denotes the natural logarithm).

The posterior-to-posterior semantics of differential privacy (as implied by the work of [KS14, NSW+17]) are as follows. Let us suppose Respondent A's information is r_t and is reported to the Census Bureau. Consider the privacy-preserving baseline in which Respondent A's data is deleted from the database and a default record r_f is added to the database to maintain the total.[9] This counterfactual corresponds to a hypothetical scenario in which all of the information in Respondent A's record is treated as private in the sense that it is unused in statistics published by the mechanism. Let D_{-1} be the database that results when the response for Respondent A is removed from the data. If the Census Bureau keeps the true value r_t then the collected data used by the Census Bureau are $D_{-1} \cup \{r_t\}$. If the Census Bureau deletes and replaces the record, then the collected data are $D_{-1} \cup \{r_f\}$. An adversary has some uncertainty about the collected data and so views it as a random variable $\mathcal{V}$. Suppose the adversary has an arbitrary prior θ about $\mathcal{V}$ (again, incorporating side information about individuals in the data along with general statistical information about the population) and is trying to make an inference about Respondent A's record, which, for the adversary is a random variable R. In our model, the attacker is allowed to know if Respondent A submitted the true record (whose value is unknown to the attacker) or submitted a default record (whose value would be known to the attacker). The Census Bureau runs a differentially private mechanism M on the data and publishes the result ω. For any possible value r of R, we can consider the ratio of the attacker's inference in both of these scenarios:

$$\begin{aligned}
&\frac{P_\theta(R = r \mid M(\mathcal{V}) = \omega)}{P_\theta(R = r \mid M(\mathcal{V}_{-1} \cup \{r_f\}) = \omega)} \\
&= \frac{\sum_{D'} P_\theta(R = r, M(D') = \omega, \mathcal{V} = D') / \sum_{D*} P_\theta(M(D^*) = \omega, \mathcal{V} = D^*)}{\sum_{D'} P_\theta(R = r, M(D'_{-1} \cup \{r_f\}) = \omega, \mathcal{V} = D') / \sum_{D*} P_\theta(M(D^*_{-1} \cup \{r_f\}) = \omega, \mathcal{V} = D^*)} \\
&= \frac{\sum_{D'} P_\theta(R = r, \mathcal{V} = D') P(M(D') = \omega) / \sum_{D*} P(M(D^*) = \omega) P_\theta(\mathcal{V} = D^*)}{\sum_{D'} P_\theta(R = r, \mathcal{V} = D') P(M(D'_{-1} \cup \{r_f\}) = \omega) / \sum_{D*} P(M(D^*_{-1} \cup \{r_f\}) = \omega) P_\theta(\mathcal{V} = D^*)}. \\
&\in [e^{-2\epsilon}, e^{2\epsilon}] \quad \text{because differential privacy guarantees } P(M(D') = \omega)/P(M(D'_{-1} \cup \{r_f\}) = \omega) \in [e^{-\epsilon}, e^{\epsilon}].
\end{aligned}$$

Thus the posterior inference about Respondent A when r_t is used differs by a factor of no more than $e^{2\epsilon}$ relative to the posterior inference about Respondent A in the private counterfactual where Respondent A's record was not used at all. In short, we say that use of the true record sharpened inference by at most $e^{2\epsilon}$.

This is the same guarantee as for randomized response, but differential privacy can help obtain better statistical efficiency. For example, the number of "yes" answers in a dataset can be released by taking the true count and adding Laplace noise with scale $1/\epsilon$ [DMNS06]. The standard deviation is $\sqrt{2}/\epsilon$ no matter how many respondents there are.

It is important to note that for an attacker who is making inferences about Respondent A, the prior θ can be arbitrary, however, the prior is over the responses/true values of other people, as well as the true value of Respondent A. The prior does not cover the actual response that A provides as input to the mechanism (this is what allows us to consider the world where A inputs the true record to the world where A inputs the default record). Thus, no matter what belief an attacker may have had about A, differential privacy guarantees that the posterior belief is not very sensitive to the record actually reported by A as input to the mechanism.

[8]The modifier *bounded* is used to emphasize that the size of the input database is publicly known.

[9]Note that the value of r_f does not depend on r_t.

4. Privacy Principles

The field of statistical disclosure limitation is over 50 years old. One may ask why the focus here is on differential privacy rather than on older methods, such as data swapping, controlled rounding, etc., [WdW96]. The reason is the emergence of important criteria for statistical disclosure methods [DMNS06, McS09, KL10].

First, mechanisms M (i.e., algorithms for statistical disclosure limitation) must be accompanied by a measure ℓ of information leakage, so that $\ell(M)$ is a quantification of how much information is leaked by M. Information leakage measures must be *closed under post-processing*. This criterion can be explained as follows. Suppose M_1 is a mechanism and A is any algorithm whose domain contains the range of M_1. Let M_2 be the mechanism that, on input D, returns $A(M(D))$. Then we require $\ell(M_2) \leqslant \ell(M_1)$. In information theory, this is known as the information processing inequality. In the case of ϵ-differential privacy, the privacy-loss budget ϵ serves as a measure of information leakage.[10] It is well-known (e.g., [McS09]) that this measure of information leakage is closed under post-processing.

The next criterion for leakage measures is *composition*. It can be explained as follows. Let D be a dataset. Let M_1 and M_2 be two mechanisms and let M_3 be the mechanism that returns $M_1(D)$ and $M_2(D)$, one could release (1) $M_1(D)$ only, (2) $M_2(D)$ only, (3) or $M_3(D)$ (which is the same as releasing $M_1(D)$ and $M_2(D)$). Clearly, the latter case should be considered more disclosive as the outputs of M_1 and M_2 together can be used to reason about confidential information in their inputs. The composition property requires that their measured information leakage be subadditive: $\ell(M_3) \leqslant \ell(M_1) + \ell(M_2)$. This allows information leakage to be treated like a monetary cost: the cost of releasing $M_1(D)$ and $M_2(D)$ together is at most the sum of their individual costs. Composition allows statistical agencies to develop statistical disclosure limitation algorithms from smaller pieces—if the total allowable information leakage is, say, at most 3, then one set of tabulations can be produced with a mechanism M_1 whose privacy leakage is measured as 1 and another set of tabulations can be produced with a mechanism M_2 whose privacy leakage is measured as 2. The combined release of both tabulations therefore satisfies the desired bound of 3.

Many statistical disclosure limitation methods do not satisfy composition. For example, two independent releases of information using k-anonymity can exactly reveal many records in the original data [GKS08].

The privacy-loss budget ϵ from differential privacy is one of the few known leakage measures that is both closed under composition and closed under post-processing.

Next, one may ask why the the focus here is on comparing posterior distributions rather than on formulating a leakage measure that compares an attacker's prior distribution to the posterior distribution after seeing the output $M(D)$ of a privacy mechanism M. The main reason is that acceptance of prior-to-posterior semantics relies on a consensus about the set of priors that is reasonable. Such a consensus is unlikely to emerge with census or other personal data. In particular, the disclosure limitation priors cannot be obtained from the input data D because such a process can lead to overfitting—tailoring the mechanism to the observed data rather than the posited sample space that generated those data.

5. Composition: where is the leakage coming from?

We now return to the problem of quantifying information leakage of a differentially private mechanism in the presence of invariants.

Consider the case of two algorithms Q and M. Algorithm Q operates on the collected data and outputs the values I of the invariants. Recall that invariants are a set of statistical tabulations that are released exactly as calculated from D, without formal privacy protections. Assume the algorithm M satisfies bounded ϵ-differential privacy and produces an output ω.

5.1. **Generally Unquantifiable Aspect of the Privacy Loss.** In the situation considered in this section, when the invariants are released there is an unavoidable initial loss of privacy. This loss generally cannot be quantified without making assumptions about an attacker's prior. For example, if an attacker knows *a priori* that block B contains at most one person and that Respondent A is the only person who could live there, then publishing the invariants reveals whether Respondent A lives in that block or not (depending on the published value of the population in that block). This attacker would be able to succeed, but other attackers

[10] Specifically, ϵ is equal to $\sup_{\omega, D_1, D_2} \log \frac{P(M(D_1)=\omega)}{P(M(D_2)=\omega)}$, where the supremum ranges over all pairs D_1, D_2 that differ on the value of one record.

with more uncertainty might not. Thus, the conclusions an attacker can make can greatly vary, depending on the priors used. There is very little consensus on which priors are reasonable to defend against. As a result, we view the initial loss of privacy due to release of invariants to be generally unquantifiable.

Applying posterior-to-posterior semantics to analyze the loss due to invariants will also not help quantify the privacy loss. To see why, consider the actual release, in which the record of Respondent A was used in the computation of the invariants and the private counterfactual world in which Respondent A reported a default record r_f (whose reported block is heavily populated). In posterior-to-posterior semantics, we assume that the attacker knows whether the true record was used or not. As the previous discussion showed, if the true record is used, the attacker may deduce the location of A; if the default record is used, the attacker may still have some uncertainty about the location of Respondent A. Thus the attacker would have significantly different inferences between the actual release and the private counterfactual (meaning privacy leakage was not bounded).

One could try to weaken this model by not allowing the attacker to know if the real or fake record was used, but this would require even more assumptions about an attacker's prior and beliefs on whether Respondent A would report the true or fake record.

5.2. **Privacy Loss Due to the Mechanism.** After invariants are published, the Census Bureau would then run a bounded ϵ-differentially private algorithm M on the data D and release $M(D)$. The combination of these two data releases will leak more information, but we can quantify how much of that extra leakage is due to M.

Consider the private counterfactual world in which the entire true data is used to compute the invariants (i.e., $Q(D)$ is released). However, before running the differentially private algorithm M, the record of Respondent A is deleted and a new record with a default value r_f is added to the database. Call this new database $D_{-1} \cup \{r_f\}$. Then M is run on this database instead of the original one and the result $M(D_{-1} \cup \{r_f\})$ is published. In this counterfactual world, we allow the attacker to know that, after the invariants were computed, the true record of Respondent A was replaced with a specific record r_f. Thus the choice of r_f must be done independently of the true record of A. In order to quantify the privacy loss solely due to the mechanism M, the natural posterior-to-posterior comparison is:

(1) *Actual Release*: Run Q(D) and $M(D)$ (which is the same as $M(D_{-1} \cup \{r_t\})$.
(2) *Private Counterfactual*: Run Q(D) and $M(D_{-1} \cup \{r_f\})$

Following the same steps as in Section 3, we get the following bound on the posterior-posterior ratio:

$$\frac{P_\theta(R = r \mid \mathrm{Q}(\mathcal{V}) = I, M(\mathcal{V}) = \omega)}{P_\theta(R = r \mid \mathrm{Q}(\mathcal{V}) = I, M(\mathcal{V}_{-1} \cup \{r_f\}) = \omega)} \in [e^{-2\epsilon}, e^{2\epsilon}] \tag{1}$$

This bound can be interpreted as follows: in the private counterfactual (which corresponds to the denominator), the differentially private mechanism is not responsible for leakage about person A because it did not use person A's record in the computation. In the actual release situation, algorithm M computes on the true data and so M has some responsibility for information leakage. However, by Equation 1, conditioning on the release of invariants I, the leakage due to the mechanism M is at most $e^{2\epsilon}$. Any further leakage that allows more accurate reconstruction of R is thus due solely to the leakage amplification caused by the invariants.

To illustrate this amplification in a simple way, consider the following artificial example adapted from [KM11]. Suppose each respondent provides an integer from the range $[0, k]$. Thus the data D consist of n integers $R_1, \ldots, R_n$. Consider the following invariants: $R_1 + R_2$, $R_2 + R_3$, $R_3 + R_4, \ldots, R_{n-1} + R_n$. There are $n - 1$ linear equations on n unknowns, so that knowing R_j for any j means all of the other R_i can be determined exactly.[11]

An example of a mechanism M that satisfies bounded ϵ-differential privacy is one that adds independent Laplace random variables Z_i with scale k/ϵ (and, thus, variance $2k^2/\epsilon^2$) to each R_i [DMNS06].

Example 1. *Had there been no invariants, then the only estimate of R_1 would be the output $R_1 + Z_1$. The mean of this estimate is R_1 (i.e., it is unbiased) and its variance is $2k^2/\epsilon^2$ (the variance of the Laplace random variable Z_1).*

[11] Note that in some situations, the invariants can be even more disclosive, for example if all of values of $R_i + R_{i+1}$ are equal to 0, then the invariants reveal that everyone's record was 0.

Example 2. *However, the invariants do exist and they do affect information leakage. Consider the actual case where M uses the true value of R_1. Because of the invariants there are now many ways to get independent estimates of R_1.*

- *The noisy estimate $R_1 + Z_1$ provides an estimate of R_1 with variance $2k^2/\epsilon^2$*
- *The noisy estimate $R_2 + Z_2$ provides another independent noisy estimate of R_1 (obtained by subtracting $R_2 + Z_2$ from the invariant $R_1 + R_2$) with variance $2k^2/\epsilon^2$.*
- *In a similar manner, each $R_j + Z_j$ provides yet another independent noisy estimate of R_1 with variance $2k^2/\epsilon^2$*

These n noisy estimates can be averaged to obtain a new estimate of R_1 that has variance (i.e. squared error) equal to $2k^2/(n\epsilon^2)$. Note that use of the constraints $0 \leqslant R_i \leqslant k$ has the potential of reducing variance even more. For example, if we know that $R_1 + R_2 = 1$, then R_1 is either 0 or 1.

Example 3. *Now suppose R_1 is changed to k before running M. Then M does not use the true value of R_1 in its computation. Now we have the following estimates:*

- *The noisy estimate $k + Z_1$ provides an (inaccurate) estimate of R_1 with squared error at most $k^2 + 2k^2/\epsilon^2$ (which would be the case if the true value was $R_1 = 0$). Call this estimate Y_1.*
- *The noisy estimate $R_2 + Z_2$ provides an independent unbiased noisy estimate of R_1 (obtained by subtracting $R_2 + Z_2$ from the invariant $R_1 + R_2$) with variance $2k^2/\epsilon^2$. Call this estimate Y_2.*
- *In a similar manner, each $R_j + Z_j$ provides yet another independent unbiased noisy estimate of R_1 with variance $2k^2/\epsilon^2$. We use Y_j to refer to the resulting estimates.*

These n noisy estimates can be averaged to obtain a new estimate of R_1 that has squared error equal to

$$E\left[\left(R_1 - \frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2\right] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n}(R_1 - Y_i)\right)^2\right] = \frac{E[(R_1 - Y_1)^2] + \sum_{i=2}^{n} E[(R_i - Y_i)^2]}{n^2}$$
$$\leqslant \frac{k^2 + 2k^2/\epsilon^2 + (n-1)2k^2/\epsilon^2}{n^2} = \frac{k^2}{n^2} + 2k^2/(n\epsilon^2)$$

Thus feeding M an incorrect value can increase the variance by at most k^2/n^2. When n is large, this increase in variance is negligible and the overall squared error of the estimate of R is small regardless of whether M used the real value of R_1 or not (and further refinements to this inference, such as using the constraints that $0 \leqslant R_i \leqslant k$ would serve to reduce the variance even more). Hence the invariants greatly amplify the information leakage.

The reason for this sharp inference is that the invariants in this example induced a strong dependence between all of the records. Thus while the resulting privacy breach is mostly due to the invariants (rather than M), the possibility of reconstruction cannot be ignored and it is desirable to have semantics for the amplified privacy leakage caused by Q possibly interacting poorly with M.

6. Tools for Analyzing Leakage Guarantees

In a statistical disclosure limitation system that publishes invariants and the output of a differentially private algorithm, both of which are computed on the same data, the privacy leakage can be decomposed into leakage of three types:

(1) The initial loss of privacy due to publication of invariants. As discussed in Section 5, this loss may not be quantifiable without making assumptions about the prior of an attacker.
(2) The loss of privacy due to the mechanism M. As discussed in Section 5, this loss can be quantified using the comparison of inferences about a person's data in two hypothetical situations: one in which person A's data is used in the computation of M versus one in which person A's data was not used in the computation of M. This privacy loss only depends on the privacy-loss budget ϵ.
(3) After the invariants have been published, there is a further loss of privacy due to the composition of the invariant release Q and the mechanism M. This additional privacy loss is a result of dependence between information released by Q and M. It consists of loss attributed to M and the amplification of this loss due to Q (as illustrated in Examples 2 and 3 in Section 5).

In this section, we develop the tools and concepts that will help provide semantics for the privacy loss in Item (3).

To analyze this type of leakage, we need to consider the following general setting of data release:

- $\mathrm{Q}(D)$ is run to produce the invariants I prior to the execution of the bounded ϵ-differentially private algorithm M.
- The input of the algorithm M is a vector I of numbers and a dataset D and the output ω of $M(I, D)$ must satisfy $\mathrm{Q}(\omega) = I$, because we assume ω contains sufficient information to uniquely determine the invariants. That is, from ω we can always recover the first input to M.
- M satisfies bounded ϵ-differential privacy with respect to D—that is, for any fixed I and for all pairs (D, D') that differ on the value of one record, and all ω, $P(M(I, D) = \omega) \leqslant e^{\epsilon} P(M(I, D') = \omega)$.[12]

Suppose I and ω were published. The inferences that are possible depend on how they were generated. Thus a first attempt to quantify the leakage in Item (3) would be to compare posterior probabilities for the following two scenarios which cause I and ω to be released:

(1) *Actual Release*: $\mathrm{Q}(D) = I, M(I, D) = \omega$—all algorithms operate on the original data.
(2) *Private Counterfactual*: $\mathrm{Q}(D_{-1} \cup \{r_f\}) = I', M(I', D_{-1} \cup \{r_f\}) = \omega$—all algorithms operate on the modified data that has a fake record for Respondent A.

However, note that unless $I = I'$, one of the two events must occur with zero probability because of the requirement that the first input to M must be recoverable from ω. As a result, if $I \neq I'$, no meaningful bound on the ratio of the posteriors can be provided (as in the discussion of Section 5 about the initial leakage caused by invariants). Thus we must develop more nuanced models of the private counterfactual that enforce $I = I'$.

To do so, we use the notion of a *modification strategy* [BGKS13]. Intuitively, one can think of a modification strategy as a procedure that scrubs information from a dataset about a person but preserves the invariants (i.e., delete record r_t and add record r_f to the dataset where r_f is chosen such that $D_{-1} \cup r_f$ satisfies the invariants I). We define modification strategies as follows.

Definition 2 (Modification Strategy)**.** A modification strategy ϕ_A is a (possibly randomized) modification of a database D that preserves the invariants $Q(D) = I$. The modified database is denoted as $\phi_A(D_{-1}, I)$ and ϕ_A satisfies the following conditions:

(1) D and $\phi_A(D_{-1}, I)$ have the same number of records, and the only difference between them is a single record: Respondent A's record, r_t, versus the replacement record, r_f.
(2) $\mathrm{Q}(\phi_A(D_{-1}, I)) = \mathrm{Q}(D)$ for all D—the modification strategy maintains consistency with invariants.

Note that the inputs to the modification strategy ϕ_A are D_{-1} and I, so ϕ_A receives no information about the reported record of Respondent A except for information necessary to compute the invariant I.

For example, suppose that the only invariants are that the total population and voting age population are known at each block. In that case, the comparing D_{-1} to I will completely determine the block and voting age status of Respondent A (but not the specific age, or reported race, etc.). So ϕ_A will have access to information about the reported voting age status and block, but will not have access to any other reported information from Respondent A. Under the modification strategy, it is as if Respondent A only provided information about block and voting-age status and opted not to provide any additional information.

Under the posterior-to-posterior approach we compare the actual release to a counterfactual in which the modification strategy is used. With this idea, we consider the following two scenarios:

(1) *Actual Release*: $Q(D) = I, M(I, D) = \omega$.
(2) *Private Counterfactual*: ϕ_A is used everywhere, $Q(\phi_A(D_{-1}, I)) = I, M(I, \phi_A(D_{-1}, I)) = \omega$.

The privacy in the counterfactual is conditional on the invariants: no information from Respondent A is used beyond what is necessary to compute the invariants. This is a relative guarantee: the leakage due to invariants is generally not quantifiable, but any leakage beyond that is protected by the modification strategy. For example, if the invariants released are highly disclosive about Respondent A, then this respondent does not have much privacy to begin with. An extreme case is if the only record that could be added to D_{-1} in

[12]Note that from the point of view of M, I is just a constraint on the output ω such that $Q(\omega) = I$. Privacy loss amplification occurs when $Q(D) = I$, that is when I is the actual invariant of the data and not just a set of numbers specified *a priori*. This notation makes privacy loss amplification easier to see, as the full algorithm (which uses the true invariants) is $M(Q(D), D)$.

order to satisfy the invariants I is Respondent A's true record r_t (see Examples 2 and 3), then the invariants are highly disclosive and there is nothing left for the modification strategy to protect about respondent A. However, if the invariants are less disclosive—for example, if at least one variable in the data is not involved in the computation of the invariants—then there always exists a non-trivial modification strategy that can scrub information about these variable(s).

The posterior-to-posterior semantics are now used to analyze the degree to which a modification strategy will affect an attacker's inference:

$$\frac{P_\theta(R = r \mid \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(R = r \mid \mathrm{Q}(\phi_A(\mathcal{V}_{-1}, I)) = I, M(I, \phi_A(\mathcal{V}_{-1}, I)) = \omega)} \tag{2}$$

where $\mathcal{V}$ is the attacker's view of the data (e.g., a random variable governed by an arbitrary distribution θ).

Theorem 1. *Let $\mathcal{V}$ be a random variable representing the true database, θ be an arbitrary prior on $\mathcal{V}$ held by an arbitrary attacker, Q be an algorithm for computing invariants, M be a bounded ϵ-differentially private mechanism that takes both invariants and a database as input. Let r be a possible record for Respondent A and let ϕ_A be a modification strategy. Then for any fixed I and ω,*

$$\frac{P_\theta(R = r \mid \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(R = r \mid \mathrm{Q}(\phi_A(\mathcal{V}_{-1}, I)) = I, M(I, \phi_A(\mathcal{V}_{-1}, I)) = \omega)} \in \left[e^{-2\epsilon}, e^{2\epsilon}\right]$$

Proof.

$$\frac{P_\theta(R = r \mid \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(R = r \mid \mathrm{Q}(\phi_A(\mathcal{V}_{-1}, I)) = I, M(I, \phi_A(\mathcal{V}_{-1}, I)) = \omega)} = \frac{P_\theta(R = r \mid \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(R = r \mid \mathrm{Q}(\mathcal{V}) = I, M(I, \phi_A(\mathcal{V}_{-1}, I)) = \omega)}$$

$$= \frac{\sum_{D'} P_\theta(R = r, \mathrm{Q}(D') = I, M(I, D') = \omega, \mathcal{V} = D') / \sum_{D^*} P_\theta(\mathrm{Q}(D^*) = I, M(I, D^*) = \omega, \mathcal{V} = D^*)}{\sum_{D'} P_\theta(R = r, \mathrm{Q}(D') = I, M(I, \phi_A(D'_{-1}, I)) = \omega, \mathcal{V} = D') / \sum_{D^*} P_\theta(\mathrm{Q}(D^*) = I, M(I, \phi_A(D^*_{-1}, I)) = \omega, \mathcal{V} = D^*)}$$

$$= \frac{\sum_{D'} P_\theta(R = r, \mathrm{Q}(D') = I, \mathcal{V} = D') P(M(I, D') = \omega) / \sum_{D^*} P(M(I, D^*) = \omega) P_\theta(\mathrm{Q}(D^*) = I, \mathcal{V} = D^*)}{\sum_{D'} P_\theta(R = r, \mathrm{Q}(D') = I, \mathcal{V} = D') P(M(I, \phi_A(D'_{-1}, I)) = \omega) / \sum_{D^*} P(M(I, \phi_A(D^*_{-1}, I)) = \omega) P_\theta(\mathrm{Q}(D^*) = I, \mathcal{V} = D^*)}$$

$\in [e^{-2\epsilon}, e^{2\epsilon}]$ because $P(M(I, D') = \omega) / P(M(I, \phi_A(D'_{-1}, I)) = \omega) \in [e^{-\epsilon}, e^{\epsilon}]$ as $D', \phi_A(D'_{-1}, I)$ differ by one record.[13]

□

Theorem 1 implies that the inference an attacker can make about respondent A after the mechanism's release is at best $e^{2\epsilon}$ times better in actuality than in the private counterfactual. This brings together the two components of the privacy guarantee: (1) the counterfactual, which can be considered private relative to the initial release of invariants, and (2) for small ϵ, attacker inference will be similar in the actual release and in the counterfactual.

It is important to note that when considering information leakage about Respondent A, any side information that is not based on the *reported* value of A can be incorporated into the arbitrary prior θ—that is, this side information can be about the true values or reported values of any other individuals in the data (note that the reported value is not always the same as the true value in practice) and can also include information about the true value of A (but not the reported value). Knowledge about the reported value of A is included in the invariants, which serves to limit the set of private counterfactuals that are possible.

7. Interpreting Guarantees for Potential 2020 Census Invariants

In this section, we consider the privacy impact of the specific invariants initially under consideration by the U.S. Census Bureau for the 2018 End-to-End Census Test and the 2020 Census. Our full analyze is included for completeness, we again note the implemented invariants are documented in [Lec19]. For illustration, assume a data schema with the following variables:

(1) Geography
(2) Relationship To Householder (Including GQ status)
(3) Sex
(4) Age
(5) Hispanic or Latino Ethnicity
(6) Race

The list of invariants for consideration is as follows:

- **C1: Total population per block.**
- **C2: Voting-age population per block.**
- **C3: Number of housing units per block.**
- **C4: Number of occupied housing units per block.**[14]
- **C5: Number of group quarters facilities by type, per block.** The count of group quarters facilities by type per block includes:
  - Group quarters type (e.g. federal prison, college dormitory).
  - Single-sex institution status (e.g. female-only, male-only, unrestricted)
  - Age restrictions (e.g. minor-only, adult-only, unrestricted)

Given a choice of invariants from **C1** - **C5**, we consider the privacy implications of releasing the invariants along with the outputs of a mechanism satisfying differential privacy. We consider the posterior-to-posterior protections offered to each respondent according to Theorem 1 relative to the private counterfactual in which the respondent reports the minimal information necessary to compute the invariants.

Case 1: **C1** only

The only invariant is the total population per block. In any reconstruction, the block-level record count will always be correct. The privacy guarantee is as follows. As per Theorem 1, the inference an attacker can make about a respondent after the mechanism's release is at best $e^{2\epsilon}$ times better than if all the attributes of a respondent's record, except the block, were replaced with arbitrary values. The privacy protection for a respondent's block-level geocode is not quantified by Theorem 1.

Case 2: **C1, C2**

The invariants are the total population per block and the voting age population per block. In any reconstruction, the total population and the voting-age population counts at the block-level will always be correct. The privacy guarantee is as follows. As per Theorem 1, the inference an attacker can make about a respondent after the mechanism's release is at best $e^{2\epsilon}$ times better than if all the attributes of a respondent's record, except the block and voting age status, were replaced with arbitrary values. This means than attributes like sex, race, and ethnicity can be modified arbitrarily. However, age can only be modified as long as the voting-age status does not change. The privacy protection for a respondent's block-level geocode or voting age status is not quantified by Theorem 1.

Case 3: **C1, C3**

The invariants are the total population per block and the number of housing units per block. In any reconstruction, the total population and the housing-unit counts at the block-level will always be correct. Recall that we define housing status as whether a respondent lives in a household and is a householder, or lives in a household and is not a householder, or lives in a group quarters. The privacy guarantee is as follows. As per Theorem 1, the inference an attacker can make about a respondent after the mechanism's release is at best $e^{2\epsilon}$ times better than if all the attributes of a respondent's record, except the block and housing status, were replaced with arbitrary values and modifications to the relationship to householder attribute (from which housing status is derived) are constrained as follows.

(1) For respondents in blocks with no housing units, no record for a person in a GQ can be altered to instead be in a household.
(2) For respondents in blocks with at least one housing unit, there are no restrictions.

The privacy protection for a respondent's block-level geocode is not quantified by Theorem 1. The privacy protection for housing status for respondents living in group quarters on blocks with no housing units is also not quantified by Theorem 1.

Case 4: **C1, C4**

The invariants are the total population per block and number of occupied housing units per block. In any reconstruction, the total population and the occupied housing-unit counts at the block-level will always be

[14]Note that this is equivalent to the number of householders per block.

correct. The privacy guarantee is as follows. As per Theorem 1, the inference an attacker can make about a respondent after the mechanism's release is at best $e^{2\epsilon}$ times better than if all the attributes of a respondent's record, except the block and housing status, were replaced with arbitrary values and modifications to the relationship to householder attribute (from which housing status is derived) are constrained as follows.

(1) A non-householder respondent's record may be altered to be a GQ record (or vice versa)
(2) A GQ or non-householder record cannot be altered to a householder record.
(3) The relation to householder of a householder cannot be modified.
(4) Respondents in blocks with only GQ units cannot have their record altered to instead be in a household.

The privacy protection for a respondent's block-level geocode is not quantified by Theorem 1. The privacy protection for housing status for respondents living in group quarters on blocks with no housing units is also not quantified by Theorem 1.

Case 5: **C1, C5**

The invariants are the total population per block and the number of group quarters facilities by type per block. In any reconstruction, the total population and the group quarters' facilities counts by type at the block-level will always be correct. The privacy guarantee is as follows. As per Theorem 1, the inference an attacker can make about a respondent after the mechanism's release is at best $e^{2\epsilon}$ times better than if race and ethnicity were modified arbitrarily, while modifications to relation to householder, sex, and age satisfy the following restrictions.

(1) Vacant group quarters are not tabulated. Individuals in group quarters containing only one person cannot change their relation to householder (i.e., they cannot change their housing status). If this group quarters has restrictions on age and sex then these attributes cannot be modified as well.
(2) For respondents living in blocks containing no group quarters, the relation to householder can be modified to any value that is valid for people living in households.

The privacy protection for a respondent's block-level geocode is not quantifiable under Theorem 1. For individuals living in a group quarters containing only one person, the protections for relation to householder, age, and sex are also not quantified by Theorem 1.

Case 6: **C3, C5**

The invariants are the number of housing units and the number and type of group quarters facilities. Per Theorem 1, the privacy protections vary for different individuals.

(1) For individuals living in a group quarters that contains only one person, the inference an attacker can make about a respondent after the mechanism's release is at best $e^{2\epsilon}$ times better than if all attributes except block, relation-to-householder, age, and sex were arbitrarily modified. Sex can be arbitrarily modified if the group quarters does not contain restrictions on sex, and age can be modified in any way that is consistent with the age restrictions in the group quarters.
(2) For any other individual, the inference an attacker can make about a respondent after the mechanism's release is at best $e^{2\epsilon}$ times better than if the entire record can be modified arbitrarily subject to the following restrictions: if the reported block is changed to a block with no households, the relation to householder cannot imply that the person lives in a household. The altered age and sex must be consistent with the allowable age and sex for group quarters in that block. If the reported block is changed to a block with no group quarters, the relation to householder cannot imply the person lives in a group quarters.

The protections for block-level geocode, relation to householder, age, and sex of respondents in group quarters containing only one person are not quantified by Theorem 1.

Case 7: **C1, C3, C5**

The invariants are the population total at each block, the number of housing units, and the number and type of group quarters facilities. The privacy guarantee combines the restrictions taken from the privacy guarantees of Cases 4-6:

(1) For respondents living in blocks that contain both GQ and housing units, a non-householder respondent's record may be altered to be a GQ record (and vice versa).
(2) A GQ or non-householder record cannot be altered to be a householder record.
(3) The relation to householder of a householder cannot be modified.
(4) Respondents in blocks with only GQ units cannot have their record altered to instead be in a household.
(5) For respondents living in blocks containing no group quarters, relation to householder can be modified to any value that is valid for people living in households.
(6) For individuals living in a group quarters that contains only one person, all attributes except block, relation-to-householder, age, and sex can be arbitrarily modified. Sex can be arbitrarily modified if the group quarters does not contain restrictions on sex, and age can be modified in any way that is consistent with the age restrictions in the group quarters.

Case 8: **C1, C2, C3, C4, C5**

The invariants are the population total at each block, the voting-age population per block, the number of housing units per block, the number of occupied housing units per block, and the number and type of group quarters per block. The privacy guarantee combines the restrictions taken from each of the prior Cases 1-7. In particular, the inference an attacker can make about a respondent after the mechanism's release is at best $e^{2\epsilon}$ times better than if block and voting-age status were unmodified, race and ethnicity were modified arbitrarily, and the rest of the attributes modified as allowed by the restrictions of Cases 1-7 (in particular, individuals living in blocks with no households have the most restrictions).

8. Group Modification Strategies

Modification strategies are designed to reason about privacy for one person at a time and do so by filling in the deleted record of a respondent with values that preserved the invariants in the resulting data. Their ability to generate semantics is limited for some attributes (such as block, when block population totals are invariant).

It is possible to extend the idea of modification strategies and private counterfactuals to reason about groups of people and provide more nuanced privacy semantics. In this extension, the records of a set $\mathcal{S}$ of respondents are first removed, and then a modification strategy fills in those missing records in order to preserve the invariants. We call this a *group modification strategy* $\varphi_{\mathcal{S}}$.

Definition 3 (Group Modification Strategy)**.** Given a set $\mathcal{S}$ of respondents, a group modification strategy $\varphi_{\mathcal{S}}$ is a (possibly randomized) modification of a database D that preserves invariants $Q(D) = I$. The modified database is denoted as $\varphi_{\mathcal{S}}(D_{-\mathcal{S}}, I)$ and satisfies the following conditions:

(1) D and $\varphi_{\mathcal{S}}(D_{-\mathcal{S}}, I)$ have the same number of records and the only differences between them are up to $|\mathcal{S}|$ records, those of the respondents in $\mathcal{S}$.
(2) $Q(\varphi_{\mathcal{S}}(D_{-\mathcal{S}}, I)) = Q(D)$ for all D—the modification strategy maintains consistency with invariants.

The inputs to the modification strategy, $D_{-\mathcal{S}}$ and I, contain no information for the respondents in $\mathcal{S}$, except what is necessary to compute the invariants I. As in the single record modification strategy case, we can consider two scenarios:

(1) *Actual Release*: $Q(D) = I, M(I, D) = \omega$.
(2) *Private Counterfactual*: $\varphi_{\mathcal{S}}$ is used everywhere, $Q(\varphi_{\mathcal{S}}(D_{-\mathcal{S}}, I)) = I, M(I, \varphi_{\mathcal{S}}(D_{-\mathcal{S}}, I)) = \omega$.

In Examples 2 and 3 there is no single record modification strategy that can be interpreted as private because the only record that can be added to D_{-1} is the true record r_t (any one-record change to the database invalidates the invariants). It is, however, possible to provide some form of guarantee using a group modification strategy. In the case of Examples 2 and 3, the set $\mathcal{S}$ would include all of the respondents.

The following theorem provides posterior-to-posterior semantics that relate the privacy provided by the private counterfactual world to the privacy provided when a mechanism M satisfying bounded ϵ-differential privacy is used in the actual world.

Theorem 2. *Let $\mathcal{S} = \{A_1, A_2, \ldots, A_k\}$ be a set of k respondents. Let $\mathcal{V}$ be a random variable representing the true database, θ be an arbitrary prior on $\mathcal{V}$ held by an arbitrary attacker, Q be an algorithm for computing invariants, M be an bounded ϵ-differentially private mechanism that takes both invariants and a database as*

input. Let $\vec{r} = [r_1, \ldots, r_k]$ *be a possible record vector for Respondents* $A_1, \ldots, A_k$ *and let* $\mathbf{R}$ *be a random variable (in the attacker's view) corresponding to the record vector for those respondents. Let* $\varphi_{\mathcal{S}}$ *be a group modification strategy. Then for any fixed* I *and* ω*,*

$$\frac{P_\theta(\mathbf{R} = \vec{r} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(\mathbf{R} = \vec{r} \mid \ \mathrm{Q}(\varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)} \in \left[e^{-2|\mathcal{S}|\epsilon}, e^{2|\mathcal{S}|\epsilon}\right]$$

Proof.

$$\frac{P_\theta(\mathbf{R} = \vec{r} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(\mathbf{R} = \vec{r} \mid \ \mathrm{Q}(\varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)} = \frac{P_\theta(\mathbf{R} = \vec{r} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(\mathbf{R} = \vec{r} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)}$$

$$= \frac{\sum_{D'} P_\theta(\mathbf{R} = \vec{r}, \mathrm{Q}(D') = I, M(I, D') = \omega, \mathcal{V} = D') / \sum_{D*} P_\theta(\mathrm{Q}(D^*) = I, M(I, D^*) = \omega, \mathcal{V} = D^*)}{\sum_{D'} P_\theta(\mathbf{R} = \vec{r}, \mathrm{Q}(D') = I, M(I, \varphi_{\mathcal{S}}(D'_{-\mathcal{S}}, I)) = \omega, \mathcal{V} = D') / \sum_{D*} P_\theta(\mathrm{Q}(D^*) = I, M(I, \varphi_{\mathcal{S}}(D^*_{-\mathcal{S}}, I)) = \omega, \mathcal{V} = D^*)}$$

$$= \frac{\sum_{D'} P_\theta(\mathbf{R} = \vec{r}, \mathrm{Q}(D') = I, \mathcal{V} = D') P(M(I, D') = \omega) / \sum_{D*} P(M(I, D^*) = \omega) P_\theta(\mathrm{Q}(D^*) = I, \mathcal{V} = D^*)}{\sum_{D'} P_\theta(\mathbf{R} = \vec{r}, \mathrm{Q}(D') = I, \mathcal{V} = D') P(M(I, \varphi_{\mathcal{S}}(D'_{-\mathcal{S}}, I)) = \omega) / \sum_{D*} P(M(I, \varphi_{\mathcal{S}}(D^*_{-\mathcal{S}}, I)) = \omega) P_\theta(\mathrm{Q}(D^*) = I, \mathcal{V} = D^*)}$$

$$\in [e^{-2|\mathcal{S}|\epsilon}, e^{2|\mathcal{S}|\epsilon}]$$

because $P(M(I, D') = \omega) / P(M(I, \varphi_{\mathcal{S}}(D'_{-\mathcal{S}}, I)) = \omega) \in [e^{-\epsilon|\mathcal{S}|}, e^{|\mathcal{S}|\epsilon}]$ as D' and $\varphi_{\mathcal{S}}(D'_{-\mathcal{S}}, I)$ differ by at most $|\mathcal{S}|$ records.[15] □

This leads to a similar interpretation of privacy guarantees for the group of individuals $\mathcal{S}$ to that of Theorem 1. In a private counterfactual world, a group of people report nothing except their identities and the group statistics that are needed to calculate the invariants. For example, if the invariants are the population total of each block and the voting age population total of each block, the group reports how many people in the group are in each block. It also reports how many people in the group are voting age in each block. Note that now the following information is missing from the reported records:

- All attributes not involved in the computation of invariants.
- The specific assignment of attribute (e.g., location) to specific members of the group.

If there are many possible different assignments of attributes to group members, then the counterfactual world may be considered private for the people in the group conditional on the invariants (i.e., relative to the initial privacy loss due to invariants). As before, the privacy leakage due to the invariants in the counterfactual world is generally not quantifiable, but any leakage beyond that is protected by the modification strategy. Theorem 2 implies that the inference an attacker can make about respondents in $\mathcal{S}$ after the mechanism's release is at best $e^{2|\mathcal{S}|\epsilon}$ times better in actuality than in the private counterfactual. This establishes the two components of the privacy guarantee: (1) the counterfactual is private relative to the initial release of invariants and (2) for small ϵ and $|\mathcal{S}|$, attacker inference will be similar in the actual release and in the counterfactual.

In addition to the group privacy guarantee conditional on the invariants, the group modification strategy can also be used to offer a more nuanced protection of invariant attributes for a respondent than the single record modification strategy. Recall that in the private counterfactual associated with the single record modification strategy, Respondent A still had to provide information about her invariant attributes in order to calculate I. By considering the differences between the invariant totals in D_{-A} and I, it would be possible to recover Respondent A's invariant attributes. As a result, no statement could be made about the protection of Respondent A's invariant attributes in the counterfactual world. However with a group modification strategy, Respondent A's invariant attributes are only provided as an aggregate along with the invariant attributes of the other respondents in $\mathcal{S}$. In this way, Respondent A's invariant attributes can be 'hidden' amongst the groups' aggregate invariant attributes, assuming the group's members have invariant attributes different from Respondent A's attributes, so that there are many possible assignments of invariant attributes to Respondent A and to the rest of the group while still maintaining the group statistics. In order for this to be meaningful, Respondent A's invariant attributes must not be able to be reconstructed from $D_{-\mathcal{S}}$ and I. This protection offered to a respondent's invariant attributes in the counterfactual is different from that offered to non-invariant attributes in the sense that the invariant attribute is still used in the mechanism.

[15] The proof reflects the deterministic case. For randomized $\varphi_{\mathcal{S}}$, we just condition on the randomness of $\varphi_{\mathcal{S}}$ until the end, and at the very end take the expectation with respect to the randomness in $\varphi_{\mathcal{S}}$.

However, since it can be hidden amongst the aggregate invariants of the group, the inference of some attackers in the counterfactual world about Respondent A's invariant attributes would be weakened, and therefore the counterfactual world is in a sense that is difficult to specify precisely (because its efficacy depends both on the particular true database and on attacker priors) more private for these attributes than with single record modification strategies. Although the counterfactual world associated with a group modification strategy can therefore be argued to be more private in a sense than that associated with a single record modification strategy, this advantage is offset by a loosening of the posterior-to-posterior bound relative to the actual release ($e^{2|\mathcal{S}|\epsilon}$ vs. $e^{2\epsilon}$).

In order to illustrate that it is necessary for invariant attributes of members in $\mathcal{S}$'s to differ from one another in order to provide additional privacy protections for those attributes in the counterfactual world, we exhibit the following two examples.

Example 4. *Suppose the only invariant is the block level population total. In this case, all members of $\mathcal{S}$ cannot belong to the same block. Otherwise $\varphi_{\mathcal{S}}(D_{-\mathcal{S}}, I)$ will fill in the correct block as there is no other possible choice to satisfy the invariant.*

Example 5. *Suppose the invariants are voting age population and total population in each block. If $|\mathcal{S}| = 2$ then both people must come from different blocks. For the same reason, they must have different voting age status. Even then, if they come from different blocks, say B_1 and B_2, then the association between block and voting age status will be revealed. For example if $D_{-\mathcal{S}}$ has one fewer person and one fewer voting age person than the invariant I for block B_1, then the person in block B_1 must be the voter. Hence $|\mathcal{S}|$ will need to be at least 4. For example two people could be in the same block B_1 but differ in voting age status and two other people could be in the same block B_2 and differ in their voting age status. By comparing $D_{-\mathcal{S}}$ to I, it is possible to know that there is 1 voting age person and 1 non-voting age person in B_1, and similarly for B_2. However, the association between id and any combination of valid block (i.e., B_1 or B_2) and voting age status is removed from the reported information.*

Thus, the group $\mathcal{S}$ has to be large enough so that for Respondent A (contained in that group), multiple assignment of attributes can be possible. The private counterfactual will then omit the specific association of attributes to the identity of Respondent A (this association might still be inferable, but not because of what the group reported). Theorem 2 then guarantees that in the actual world, any posterior inference about the specific association of attributes to Respondent A (or, in general any property of the group), will change at most by a factor of $e^{2|\mathcal{S}|\epsilon}$.

9. Tools for Analyzing Leakage Guarantees as Odds Ratios

Recall that group modification strategies are designed to reason about how well an attacker can distinguish between alternatives (e.g., is Respondent A a voting age person in Block B_1 vs. non-voting age person in B_1 vs. voting age person in B_2 vs. non-voting age person in B_2). In this section we extend our privacy semantics with the use of group modification strategies to study the ability of an attacker to discriminate between alternatives.

For each person, we specify a set of secret pairs $\mathbb{S}_{\text{pairs}} = \{(\sigma_{a_1}, \sigma_{b_1}), \ldots, (\sigma_{a_\ell}, \sigma_{b_\ell})\}$ indicating that we are interested protecting against an attacker using the output of M to improve his inference about how likely it is that σ_{a_i} is true about a person's record r compared to σ_{b_i} being true. We express those statements mathematically as $\sigma_{a_i}(r) = \text{True}$ and $\sigma_{b_i}(r) = \text{True}$, respectively. Note that, for each i, σ_{a_i} and σ_{b_i} must be mutually exclusive.

To measure the relative likelihood of one secret compared to another, we treat the record as a random variable R (since the attacker does not know it) and we calculate the *odds*. The *odds* are the probability that $\sigma_{a_i}(R) = \text{True}$ given an attacker's prior θ, the output of M, and the output of Q divided by the probability that $\sigma_{b_i}(R) = \text{True}$ given an attacker's prior θ, the output of M, and the output of Q:

$$\frac{P_\theta(\sigma_{a_i}(R) = \text{True} \mid \text{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(\sigma_{b_i}(R) = \text{True} \mid \text{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}$$

We compare these odds to the odds assuming a private counterfactual in which there is a group $\mathcal{S}$ of people whose information is removed and who only report the group statistics that are necessary to maintain the invariants. We employ a group modification strategy that replaces the deleted records with values that

satisfy the invariants. A properly chosen group will mean that there are several possible assignments of attributes to the record of Respondent A and in the private counterfactual no information about any of these possible assignments are reported. In this private counterfactual world, the odds are:

$$\frac{P_\theta(\sigma_{a_i}(R) = \text{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)}{P_\theta(\sigma_{b_i}(R) = \text{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)}$$

The privacy leakage is then the following odds ratio and represents the improvement in inference due to using true records rather than the modification strategy which scrubs records:

$$\begin{aligned}
&\frac{P_\theta(\sigma_{a_i}(R) = \text{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(\sigma_{b_i}(R) = \text{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)} \Big/ \frac{P_\theta(\sigma_{a_i}(R) = \text{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)}{P_\theta(\sigma_{b_i}(R) = \text{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)} \\
&= \frac{P_\theta(\sigma_{a_i}(R) = \text{True} \ , \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(\sigma_{b_i}(R) = \text{True} \ , \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)} \Big/ \frac{P_\theta(\sigma_{a_i}(R) = \text{True} \ , \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)}{P_\theta(\sigma_{b_i}(R) = \text{True} \ , \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_{\mathcal{S}}(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)} \\
&= \frac{\sum_D P_\theta(\sigma_{a_i}(R) = \text{True} \ , \ \mathrm{Q}(\mathcal{V}) = I, \mathcal{V} = D) P(M(I, D) = \omega)}{\sum_D P_\theta(\sigma_{b_i}(R) = \text{True} \ , \ \mathrm{Q}(\mathcal{V}) = I, \mathcal{V} = D) P(M(I, D) = \omega)} \Big/ \frac{\sum_D P_\theta(\sigma_{a_i}(R) = \text{True} \ , \ \mathrm{Q}(\mathcal{V}) = I, \mathcal{V} = D) P(M(I, \varphi_{\mathcal{S}}(D_{-\mathcal{S}}, I)) = \omega)}{\sum_D P_\theta(\sigma_{b_i}(R) = \text{True} \ , \ \mathrm{Q}(\mathcal{V}) = I, \mathcal{V} = D) P(M(I, \varphi_{\mathcal{S}}(D_{-\mathcal{S}}, I)) = \omega)} \\
&\in [e^{-2\epsilon|\mathcal{S}|}, e^{2\epsilon|\mathcal{S}|}]
\end{aligned}$$

when M satisfies bounded ϵ-differential privacy. This result is summarized as follows:

Theorem 3. *Let $\mathcal{V}$ be a random variable representing the true database, θ be an arbitrary prior on $\mathcal{V}$ held by an attacker, Q be an algorithm for computing invariants, M be an bounded ϵ-differentially private mechanism that takes both invariants and a database as input. Let R be the random variable (in the view of the attacker) corresponding to the record for Respondent A. Let $\mathcal{S}$ be a group of people and let $\varphi_{\mathcal{S}}$ be a group modification strategy. Then for any fixed I and ω and any secret pair $(\sigma_{a_i}, \sigma_{b_i})$,*

$$\frac{P_\theta(\sigma_{a_i}(R) = \textit{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)}{P_\theta(\sigma_{b_i}(R) = \textit{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \mathcal{V}) = \omega)} \Big/ \frac{P_\theta(\sigma_{a_i}(R) = \textit{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_S(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)}{P_\theta(\sigma_{b_i}(R) = \textit{True} \mid \ \mathrm{Q}(\mathcal{V}) = I, M(I, \varphi_S(\mathcal{V}_{-\mathcal{S}}, I)) = \omega)} \in [e^{-2\epsilon|\mathcal{S}|}, e^{2\epsilon|\mathcal{S}|}]$$

Note that Theorem 2 can be used directly to provide odds ratio bounds as well, but those bounds would be worse (i.e., $e^{4|\mathcal{S}|}$ instead of $e^{2|\mathcal{S}|}$).

10. Guarantees for Attributes Involved in Computation of Census Invariants

In Section 7, we considered the privacy impact of invariants under Theorem 1. In short, attributes involved in the computation of invariants were not given quantifiable protection but the rest of the attributes had quantifiable protections. In this section, we use Theorem 2 and Theorem 3 to provide more nuanced guarantees for these attributes. The schema and invariants are as defined in Section 7.

Case 1: **C1** only.

As per Theorem 3, the inference an attacker can make about a respondent's block after the mechanism's release is at best $e^{4\epsilon}$ times better than if the respondent's block was swapped with the block of another respondent.

Case 2: **C1, C2**

As per Theorem 3, the inference an attacker can make about a respondent's block and voting age status after the mechanism's release is at best $e^{8\epsilon}$ times better than if the voting age status and block of the respondent and 3 other individuals were arbitrarily reassigned. More specifically, for any group of 4 individuals consisting of one voting age and one non-voting age person in one block and one voting age and one non-voting age person in a second block, the assignment of block and voting age status to each individual is protected by an inference bound of $e^{8\epsilon}$ compared to a private counterfactual in which those attributes in their records were randomly re-assigned among the 4 individuals.

Case 3: **C1, C3**

Under the basic analysis of Section 7, there were no quantifiable protections of housing status and block, and in some cases age and sex. However, with Theorem 3, the inference an attacker can make about a respondent's record is at most $e^{4\epsilon}$ times better than if the block and housing status of a respondent were

swapped with that of another individual, and the rest of the record was altered arbitrarily.

Case 4: **C1, C4**

The privacy guarantees provided by Theorem 3 are the same as in Case 3.

Case 5: **C1, C5**

Under Theorem 3, the inference an attacker can make about a respondent's record is at most $e^{4\epsilon}$ times better than if the block and housing status of a respondent were swapped with that of another individual, and the other values in this pair of records were modified arbitrarily subject to edit rules. For example, if the respondent is from a block containing only female-only dormitories, then after swapping blocks with another individual, that individual's reported sex would be changed to female by edit rules. However, the assignment of which individual belongs to which block and has which sex is protected with an inference bound of at most $e^{4\epsilon}$ compared to a random re-assignment of blocks to the pair of individuals.

Case 6: **C3, C5**

We consider two situations:

Option 1: Vacant group quarters are not tabulated, and so the invariants reveal a minimum on the population in each block, and Theorem 3 is needed. For individuals living in group quarters containing only one person, as per Theorem 3, inference about housing status is protected by a bound of at most $e^{4\epsilon}$ than if the housing status of the respondent's record was swapped with that of another record and both records were modified arbitrarily (subject to edit rules as in case 5). For individuals not living in a single-person group quarters, as per Theorem 3, inference about any part of the response is protected by an inference bound of at most $e^{2\epsilon}$ than if the entire record was arbitrarily modified.

Option 2: If vacant group quarters were also tabulated, then the invariants leak no information about any response. The data release would enjoy the full protection of ϵ-differential privacy. Inference about any respondent's record is at most $e^{2\epsilon}$ better than if the entire record was arbitrarily modified, with the exception that it cannot be modified to a person in a household in a block with no housing units, and it cannot be modified to a group quarters person in a block having no group quarters.

Case 7: **C1, C3, C5**

The analysis is similar to Cases 3 and 5. Under Theorem 3, the inference an attacker can make about a respondent's record is at most $e^{4\epsilon}$ times better than if the block of a respondent was swapped with the block of another individual, and the other values in this pair of records were modified arbitrarily subject to edit rules (as in Case 5).

Case 8: **C1, C2, C3, C4, C5**

When all five invariants are used, the privacy guarantees are most complicated and have the weakest upper bound in inference. If the respondent is in any group of 8 people, with 4 from one block (having all combinations of voting age status and housing status) and 4 from another block (having all combinations of voting status and housing status), then Theorem 3 guarantees that the inference an attacker can make about a respondent's record is at most $e^{16\epsilon}$ times better than if the assignment of block, voting age status and housing status were randomly reassigned within the group.

11. Conclusions

The release of exact statistics (i.e., invariants) about a dataset causes a privacy leakage amplification—the normal privacy leakage of a statistical disclosure limitation algorithm gets amplified when its output is combined with invariants. In this paper, we extended the semantics of differential privacy to analyze this amplified privacy leakage. The analysis required the use of group modification strategies that are designed to remove information before any algorithm computes on the data. The resulting privacy guarantees relate the posterior inference about an individual's reported data compared to the private counterfactual world in which the modification strategies were used.

We applied these results to the invariants under consideration for the 2020 Census of Population and Housing. The information leakage due to the invariants is generally not quantifiable, but any subsequent leakage (e.g., when the invariants are combined with the output of a differentially private algorithm) can be quantified and separated into leakage due to disclosure limitation and leakage due to the amplification caused by the invariants.

References

[BGKS13] Raef Bassily, Adam Groce, Jonathan Katz, and Adam D. Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *FOCS*, pages 439–448, 2013.

[Bur12] U.S. Census Bureau. 2010 summary file 1 technical documentation. `https://www.census.gov/prod/cen2010/doc/sf1.pdf`, 2012.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. 2006.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2003.

[DN10] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):93–107, 2010.

[Fel72] I. P. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337):7–18, 1972.

[GKS08] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, 2008.

[KL10] Daniel Kifer and Bing-Rong Lin. Towards an axiomatization of statistical privacy and utility. In *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, 2010.

[KM11] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 193–204, New York, NY, USA, 2011. ACM.

[KM12] Daniel Kifer and Ashwin Machanavajjhala. A rigorous and customizable framework for privacy. In *ACM Symposium on Principles of Database Systems (PODS)*, 2012.

[KM14] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1):3, 2014.

[KS14] Shiva P Kasiviswanathan and Adam Smith. On the'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1):1–16, 2014.

[KS15] Shiva Prasad Kasiviswanathan and Adam Smith. On the 'semantics' of differential privacy: A bayesian formulation. 2015.

[Lec19] Phil Leclerc. Guide to the census 2018 end-to-end disclosure avoidance algorithm and implementation, 2019.

[McS09] Frank D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, 2009.

[NSW+17] Kobbi Nissim, Thomas Steinke, Alexandra Wood, Mark Bun, Marco Gaboardi, David R. O'Brien, and Salil Vadhan. Differential privacy: A primer for a non-technical audience (preliminary version). 2017.

[U.S02] U.S. Census Bureau. Census Confidentiality and Privacy 1790 to 2002. Technical report, Department of Commerce, 2002. (Cited on March 22, 2018).

[War65] S. L. Warner. Randomised response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63—69, March 1965.

[WdW96] Leon Willenborg and Ton de Waal. *Statistical Disclosure Control in Practice*. Springer-Verlag New York, 1996.