Series Name: WPS

Paper No.: 242

Issue Date: 9/4/24

Enabling Humanitarian Applications with Targeted Differential Privacy

Nitin Kohli and Joshua E. Blumenstock

CEGA

Center for Effective Global Action

Working Paper Series

Center for Effective Global Action

University of California

eScholarship
University of California

Recommended Citation:

Kohli, Nitin and Joshua E. Blumenstock (2024): Enabling Humanitarian Applications with Targeted Differential Privacy. CEGA Working Paper Series No. WPS-242. Center for Effective Global Action. University of California, Berkeley. Text. https://doi.org/10.26085/C3359B

Enabling Humanitarian Applications with Targeted Differential Privacy

Nitin Kohli[1] and Joshua E. Blumenstock[1,2,*]

[1]Center for Effective Global Action, UC Berkeley
[2]School of Information, UC Berkeley
[*]Corresponding Author: jblumenstock@berkeley.edu

Abstract

The proliferation of mobile phones in low- and middle-income countries has suddenly and dramatically increased the extent to which the world's poorest and most vulnerable populations can be observed and tracked by governments and corporations. Millions of historically "off the grid" individuals are now passively generating digital data; these data, in turn, are being used to make life-altering decisions about those individuals — including whether or not they receive government benefits, and whether they qualify for a consumer loan. This paper develops an approach to implementing algorithmic decisions based on personal data, while also providing formal privacy guarantees to data subjects. The approach adapts differential privacy to applications that require decisions about individuals, and gives decision makers granular control over the level of privacy guaranteed to data subjects. We show that stronger privacy guarantees typically come at some cost, and use data from two real-world applications — an anti-poverty program in Togo and a consumer lending platform in Nigeria — to illustrate those costs. Our empirical results quantify the tradeoff between privacy and predictive accuracy, and characterize how different privacy guarantees impact overall program effectiveness. More broadly, our results demonstrate a way for humanitarian programs to responsibly use personal data, and better equip program designers to make informed decisions about data privacy.

Introduction

Mobile phones are now close to ubiquitous in even the world's poorest nations. It is estimated that 81% of women and 87% of men in low- and middle-income countries (LMICs) own a mobile phone [1], and roughly 150 million individuals used mobile internet for the first time in 2023 [2]. With continual advances and investments in digital public infrastructure [3, 4] and network connectivity, these numbers are expected to rise [5]. Consequently, billions of historically "off the grid" individuals are now passively generating digital footprint data through the everyday use of their mobile devices.

These data, in turn, are being used by governments and corporations to make life-altering decisions about individuals in a variety of humanitarian applications. For example, digital footprint data from mobile phone networks have been successfully used in social protection and anti-poverty initiatives to identify individuals with the greatest need for humanitarian support [6]. Mobile phone metadata have likewise been used in consumer credit products across LMICs to determine loan eligibility for hundreds of millions of historically unbanked individuals [7]. In each of these settings, people's digital footprint data are combined with machine learning algorithms to *target* them, i.e., to determine whether they should be eligible for some benefit (e.g., cash transfers or loans).

However, the metadata generated when people use their mobile phones is exceptionally sensitive. Prior work has shown that mobile phone metadata can reveal an individual's sexual orientation [8, 9], religious affiliation [10], political preferences [11], and social network connections [12]. For this reason, the use and analysis of these data — particularly from socioeconomically disadvantaged populations — raises serious privacy concerns [13, 14, 15, 16].

This paper develops and tests a novel approach to making algorithmic targeting decisions based on provably private data, which provides formal and robust privacy guarantees while simultaneously enabling accurate interventions in downstream applications. We make three main contributions. Our first contribution is methodological and provides a framework for rigorously reasoning about privacy in targeting applications. We illustrate how a variant of differential privacy — which we call *targeted differential privacy* — can be used to produce provably private datasets that still allow for accurate targeting. Our second contribution is technical and provides a novel algorithm to generate privatized datasets that satisfy targeted differential privacy. Finally, we provide robust evidence on the empirical tradeoff between privacy and predictive accuracy in two real-world settings: an anti-poverty program in Togo and a consumer lending platform in Nigeria. The analysis shows how higher levels of privacy can protect against two canonical threats (singling-out attacks and attribute inference attacks), but that such privacy protections can impact overall program effectiveness. Perhaps most notably, we find that large increases in privacy can often be obtained for relatively small sacrifices in targeting accuracy. Taken together, our results illustrate how mathematical techniques can enable the responsible use of personal data in humanitarian programs, and better equip program designers to make informed decisions about the tradeoffs involved in using targeted differential privacy in practice.

A framework and algorithm for provably private targeting

Our first contribution is a framework for rigorously reasoning about data privacy in targeting applications. Underlying this framework is an interaction between a data holder (such as a mobile network operator) and a downstream policymaker or program official who wishes to make use of the data (such as a humanitarian program manager or loan officer). The program official seeks to target a policy (or product, promotion, etc.) to individuals based on eligibility criteria that are difficult or costly to observe directly, such as individual levels of consumption, deprivation, creditworthiness, or profitability. (This is often the case in humanitarian settings when, for instance, the government does not have accurate and up-to-date data on impacted populations [17]). We consider the common case in which the policymaker wishes to use a decision rule to determine eligibility based on *proxy* information, where the proxies that form the decision rule are learned from patterns observed in a subset of the population for whom eligibility data can be directly observed (i.e., a learning or training sample). In traditional anti-poverty programs, administrators use simple linear models to identify observable household characteristics (such as roofing material) that can proxy for more nuanced measures of welfare (such as food consumption) [18]; in more recent programs, sophisticated machine learning algorithms are applied to high-dimensional data from multiple sources to predict household wealth [6]. Likewise, traditional lending decisions are based on formal financial histories; more recent "alternative" credit scores use data from mobile phones and social media [19].

We study the privacy implications of this interaction when the program official wishes to base decisions on proxy data X that are owned by the data holder, where X contains sensitive information. To protect the privacy of the data, the data holder provides the program official a privacy-preserving dataset X_{priv} that is generated by some privacy-enhancing technology. The program official then uses X_{priv} as the basis for a decision rule that determines program eligibility. We refer to this data sharing interaction and prediction process as the *targeting setting*; it is depicted in Figure 1(A-F).

For most targeting settings, existing privacy-enhancing technologies (PETs) perform poorly. To illustrate, Figure 2 shows how two different programs — an anti-poverty program in Togo (Figure 2A) and a micro-lending platform in Nigeria (Figure 2B) — are impacted by common PETs. We provide details on both programs below, but at a general level both programs used machine learning, in conjunction with personal data from mobile phones, to determine an individual's eligibility for a valuable benefit. In Togo, where the benefit was an emergency cash transfer, we show results for a program in which 29% of 4.95 million total individuals were eligible. In Nigeria, where the benefit was a micro-loan, we show results for a program in which low-risk borrowers (i.e., those with a high probability of repaying the loan) are eligible. In Figure 2, the left-most blue bar indicates baseline performance before any privacy protections are provided, and the two middle red bars indicate performance using differential privacy [20, 21] and k-anonymity [22, 23]. In the anti-poverty initiative, predictive accuracy falls by $4.7\% - 9.7\%$; for a national anti-poverty program this would translate into $115K - 240K$ additional exclusion errors (i.e., true poor who were excluded from receiving benefits as a result of inaccurate targeting). In the Nigerian consumer lending application, the accuracy of the credit scoring model fell by $16.1\% - 16.4\%$; had the micro-lending platform utilized these PETs, the relative profit of the lending program would have been reduced by $430\% - 476\%$ (see *Methods*, *Case Studies*).

This stark performance loss arises from a conceptual mismatch between existing PETs and the targeting setting. Accurate targeting relies on an algorithm's ability to correctly distinguish between individuals of different types, based on observed characteristics of those individuals. However, differential privacy is designed to enable statistical learning about *groups* of individuals, while restricting "excessive" learning about any individual in the group [24] ("excessive" is made precise in *Methods*, Definition 1); this process thus obscures information about individuals that is critical to targeting. k-anonymous algorithms, by contrast, are designed to provide "protection in the crowd" by ensuring that each individual in the privatized dataset appears identical to at least $k-1$ other individuals [22]; this process thus inhibits a targeting application's ability to differentiate between individuals who have different eligibility statuses, but who are grouped together in the k-anonymous algorithm.

This motivates the definition of (B, ϵ, δ)-*targeted differential privacy* (TDP). At a conceptual level, TDP augments the definition of differential privacy by introducing an auxiliary parameter B to ensure that X_{priv} contains sufficient information to delineate between "sufficiently different" individuals, but insufficient information to delineate between "sufficiently similar" individuals (see Definition 2 in *Methods*). This enables TDP to interpolate between two extremes: as B increases, the privacy TDP provides approaches that of (ϵ, δ)-differential privacy; as B decreases, TDP privacy approaches the protection already present in X. By construction, TDP algorithms can generate private datasets that enable algorithms to delineate between sufficiently different individuals, thereby permitting accurate targeting outcomes. For more information on contextual adaptations made to the definition of differential privacy, see *Methods* (*Related Contextual Adaptations*).

We formalize the relationship between TDP and accurate targeting in Theorem 1 (see *Supplemental Information*), which has two implications for the theory and practice of targeting using privatized data. First, this theorem demonstrates an inherent limitation on the use of differential privacy in the targeting setting: as we strengthen the privacy parameters of any differentially private algorithm, we diminish the ability to correctly target individuals. In particular, this theorem shows that results for differential privacy in Figure 2 are not an accident – in general, differentially private algorithms cannot reliably generate X_{priv} that provide accurate predictions for targeting. Second, Theorem 1 provides potential values for (B, ϵ, δ) to achieve a desired level of targeting accuracy for program officials to consider as they configure TDP algorithms (see Example 1 in *Methods*). Together, the definition of TDP and Theorem 1 provide a framework that makes it possible to translate the needs of program designers into values for (B, ϵ, δ) that enable provably private and accurate targeting.

Our second contribution is a novel algorithm that satisfies the definition of TDP. This innovation is required because other common privatization strategies — beyond differential privacy and k-anonymity — are likewise inappropriate for targeting settings. For instance, we cannot simply delete individuals through subsampling [25], as this process would mechanically exclude them from eligibility for the benefit. Likewise, it is not viable to insert "fake" individuals [25, 26, 27], as targeting fake individuals reduces resources available to real individuals. Instead, Algorithm 1 provides a general-purpose method for constructing privatized datasets satisfying TDP (see *Methods, Private Projection Algorithm*). Figure 1(G)-(I) provides a conceptual schematic of the process. The algorithm adapts theory of differentially private projections [28, 29, 30] and Johnson–Lindenstrauss transforms [31, 32], and behaves as follows. The records in a original dataset X are randomly mapped to a higher-dimensional space in a manner that satisfies TDP with favorable statistical and scaling properties: as the dimensionality of this space increases, the amount of noise needed to satisfy TDP decreases. To avoid the difficulties that arise in high-dimensional machine learning, the algorithm then projects these datapoints back into the original dimensionality using a two-step process: first, we privately learn the statistical structure in X by computing a TDP covariance matrix; we then use the right singular values of this privatized covariance matrix to return the datapoints to their original space. The resulting set of datapoints (X_{priv}) satisfy TDP, and have values that approximately preserve the statistical structure of X. This is of central importance for machine learning applications, as similar records in X tend to remain similar in X_{priv}, and dissimilar records in X tend to remain dissimilar in X_{priv}.

Evaluation of the privacy-program effectiveness tradeoff

Our third set of results empirically characterize the tradeoffs induced by targeted differential privacy (TDP). We perform this analysis using data from the two real-world applications previewed in Figure 2: an anti-poverty program in Togo and a consumer lending platform in Nigeria. Broadly, we observe that TDP induces a *privacy-accuracy tradeoff* [24, 33]: as the privacy guarantee becomes stronger, the performance of the downstream application is degraded. However, considerable nuance can be found in the nature of these tradeoffs.

Since privacy can be infringed upon in a multitude of ways [34, 35], our empirical analysis quantifies the extent to which TDP provides protection against three distinct privacy threats that have been of central concern to legislators and privacy professionals. The first are *singling-out attacks*, where an adversary isolates an individual's data in X. Singling-out attacks are explicitly mentioned in Europe's flagship data privacy law, the General Data Protection Regulation (GDPR, Article 29 WP 216) [36]. The second threat we consider, also referenced in the GDPR, are *attribute inference attacks*, where an adversary infers the fields of X. The last threat we consider are *distinguishing attacks*, where an adversary distinguishes between likely and unlikely values in X. For all three attacks, we provide a privacy protection score that ranges between 0 and 1, with higher values corresponding to stronger privacy guarantees. For technical details on the attacks and scores, see *Methods* (*Privacy Attacks*).

Humanitarian aid program in Togo

We first characterize the privacy-program effectiveness tradeoff in the context of a humanitarian aid program. In 2020, the government of Togo launched the *Novissi* emergency social assistance program to provide cash transfers to needy individuals during the COVID-19 pandemic. When the program was launched in rural areas, the government did not have data to indicate which individuals had the greatest need for assistance; instead, they used a combination of machine learning and personal data. Specifically, working with researchers, they trained a machine learning model to predict each individual mobile subscriber's poverty status using mobile phone metadata obtained from the country's two mobile phone operators. While the poverty scores produced from mobile phone metadata were imperfect, they were significantly more accurate than the other targeting mechanisms available to the government during the COVID-19 pandemic [6].

While humanitarian crises may justify the use of personal data for the greater good [37, 38], there still remains an imperative to protect personal privacy in such applications [39]. We present results that compare the exclusion errors (i.e., the number of individuals who are truly poor but who are incorrectly excluded from the program) when privatized data are used in place of original data. Figure 3A illustrates the tradeoff between protection against singling-out attacks (which increase along the x-axis) and program effectiveness (which decreases in the number of exclusion errors, shown on the y-axis). The curve illustrates how the tradeoff between program effectiveness and privacy protection varies with different values of B, the key tuning parameter for our TDP algorithm. There are three points labeled on the curve: a blue star, representing the non-private status quo (i.e., what was used by the government of Togo at the time); a green star, corresponding to our algorithm when $B = 0.25$; and a red star, corresponding to traditional differential privacy. As B increases, TDP provides increasingly strong privacy protections (singling-out protection approaches 1), consistent with prior mathematical results proving that differential privacy thwarts singling-out attacks (provided the parameters ϵ and δ are sufficiently small) [40]. When $B = 0.25$, the algorithm reaches a balance between two extremes: compared to the non-private status quo, our approach incurs a modest increase in exclusion errors in exchange for substantive increases in privacy protection ($2K$ additional exclusion errors vs. 8,480% increase in singling-out protection); compared to differential privacy, our approach incurs a modest loss in privacy protection in exchange for a significant reduction in exclusion errors (12.6% decrease in singling-out protection vs. $113K$ fewer exclusion errors).

Figure 3B illustrates a similar, albeit more nuanced, tradeoff between program performance and protection against attribute inference attacks. When $B = 0.25$, our method provides a substantial increase in privacy protection over the non-private status quo (by 7,300%); compared to differential privacy, our approach incurs a reduction in protection of just 3.4%. Unlike singling-out protection, we find that increasing B does not result in a monotonic increase in attribute inference protection (as evidenced by the jagged shape of the tradeoff curve), and the attribute inference protection never reaches 1. This arises because differential privacy, as well as targeted differential privacy, use noise to mask identifiable information across individuals; this is distinct from using noise to mask statistical information across attributes, which would thwart an attribute inference attack. Similar results have been observed in the field of adversarial machine learning, where attribute inference attacks have successfully thwarted differentially private machine learning models [41, 42].

Figure 3C illustrates the tradeoff between program effectiveness and protection against distinguishing attacks. Compared to the non-private status quo, the value $B = 0.25$ increases distinguishing protection from 0 to 6.66%, at the cost of $2K$ additional exclusion errors. Differential privacy provides substantially greater protection against distinguishing attacks (83.5%), but at the considerable cost of $115K$ additional exclusion errors. These sharper tradeoffs arise from the fact that the curve in Figure 3C is approximately linear, which in turn arises from Theorem 1, since distinguishing attacks are intimately related to the protection provided by differential privacy [43]. More generally, the shape of the curve implies that it may be difficult to achieve strong protections against distinguishing attacks while also allowing for accurate targeting, since accurate targeting requires the preservation of information to distinguish between eligible and ineligible individuals.

Lastly, we note that increased privacy does not always require a reduction in program effectiveness. For instance, when $B < 0.25$, the private algorithm actually achieves marginally higher program effectiveness than when $B = 0$. This finding is consistent with results in the machine learning literature that indicate that the careful introduction of noise can — in some situations — improve the predictive accuracy of a model by reducing overfitting [44, 45, 46].

Consumer lending application in Nigeria

Our second empirical example characterizes the privacy-program effectiveness tradeoff in the context of a micro-lending platform in Nigeria. Like many other such platforms that have recently gained widespread popularity in LMICs [47], the platform we study leverages the mobile phone network to distribute and collect payments on very small loans (roughly $10). A distinguishing feature of these "digital credit" providers is that, since many of their customers do not have bank accounts or formal financial histories, they make lending decisions based on alternative credit scores that are derived from customers' digital "data footprints." The insight behind these alternative credit scores is that data passively generated by mobile phone use – such as the volume of international phone calls made by the customer – are excellent predictors of loan repayment [19]. Lenders use machine learning algorithms to detect these patterns, and offer loans to people with a high probability of repayment. While many are excited by the potential for these products to "bank the unbanked" [48], there are also privacy concerns about how personal data are being used [16].

We present results that compare the profitability of the lending program when privatized data are used in place of original data. We evaluate privacy as protection against the three attacks discussed previously, and measure profit as the revenue generated by lending decisions that result in repayment, minus the loss incurred by lending decisions that result in default (see *Methods, Case Studies*).

Broadly, we observe a similar qualitative pattern in the Nigerian lending platform as with the humanitarian aid program in Togo: there are stark tradeoffs to be made at the two extremes of privacy protection, but large gains in privacy can be obtained for modest sacrifices in profit. The tradeoff between lender profit and singling-out protection is shown in Figure 3D, with three points of interest: a blue star, representing the non-private status quo; a green star, corresponding to our algorithm when $B = 0.1$; and a red star, corresponding to differential privacy. We find that, relative to the non-private approach, our algorithm increases singling-out protection by 7,936% in exchange for a 12% reduction in relative profit; relative to the differentially private approach, our approach increases relative profit by 79%, at the expense of a 11% loss in privacy protection. Figure 3E displays the relationship between attribute inference inference and the program's profitability. Interestingly, we find that our approach has stronger privacy protection than *both* the non-private approach and the differentially private approach (an increase of 8,788% and 9% increase respectively). In particular, increasing B from 0 to 0.05 yields drastic increase attribute inference protection, whereas subsequently increasing B to 0.1 results in a negligible increase in protection. In contrast to the prior two attacks, Figure 3F displays a sharp tradeoff between distinguishing protection and the program's profitability, consistent with the results of Theorem 1 that accurate targeting necessarily requires smaller values of B.

Navigating tradeoffs in practice

While the broad takeaway from both case studies is the same — that there generally exists a tradeoff between privacy and program effectiveness, and that large gains in one can be obtained for small reductions in the other — there are nuanced differences between the cases that highlight the sort of tradeoffs that program officials may need to make in practice. Across both programs, for instance, increasing B results in a monotonic increase in protection against singling-out and distinguishing attacks. However, in both programs we find that attribute inference protection increases very rapidly up to a certain point; beyond that point, little protection is gained by increasing B. Interestingly, the protection is neither monotonic in B, nor does it approach 1 as we increase B. In fact, our results in Nigeria demonstrate that differential privacy actually exhibits *lower* attribute inference protection than targeted differential privacy when $B = 0.1$.

Taken together, these results suggest that there may be more than just a privacy-accuracy and privacy-program effectiveness tradeoff; in some situations, there may be a *privacy-privacy tradeoff*, where, for instance, increases in singling-out protection are achieved through decreases in attribute inference protection (and vice-versa). When such situations arise, data holders and program officials must collaborate to decide which types of privacy protections are most important to prioritize, subject to accuracy constraints. In the two empirical settings we study, singling-out and attribution inference protections are not as costly to provide, so they may be easier to prioritize. However, such decisions imply a different set of societal values and priorities, and are deeply contextual [49]. The decisions must account for several factors, including the contents of the specific data in question, relevant privacy laws and international humanitarian standards, and the feasibility of deploying additional privacy-monitoring systems (e.g., query logging can be used to record the computations executed on the privatized data; while imperfect [50], logging can be used to monitor downstream users for adversarial behavior and reduce privacy risk [51, 52]).

Discussion

This paper develops a framework and efficient algorithm for targeted differential privacy: a paradigm for constructing provably private datasets that still enable applications that make targeting decisions about individuals. As illustrated by the case studies in Togo and Nigeria, those privacy guarantees can impact program effectiveness. While the exact nature of that tradeoff is context-dependent, we consistently observe that large gains in privacy can be obtained for small reductions in program effectiveness.

The insights from Togo and Nigeria are likely to generalize to a wide range of settings where personal data are used to determine eligibility for policies, products, and benefits. We focus on two low- and middle-income countries (LMICs) because these are settings where, until recently, most individuals were not generating rich "digital footprint" data that could be re-analyzed by companies and policymakers. Over the past few years, however, a wide range of LMIC policies and products now rely on sensitive personal data to make high-stakes decisions — including in settings of food security, financial services, and social protection [16, 53, 54, 55]. While our focus on LMICs is thus not accidental, we expect that our framework and algorithm for targeted differential privacy could likewise be applied in high-income settings.

As we consider how privacy protections can be provided for other applications, more work is required to characterize the tradeoff between privacy and program effectiveness in each new setting. This is will require deep collaboration between data holders, program designers, and the research community. However, the technical approach we propose should not be mistaken as a substitute for the privacy protections afforded by privacy laws, policies, and practices — rather, it complements existing protection mechanisms by providing assurances about a privatized dataset's resilience to technical data privacy attacks. We hope future research will build upon our work, and generate a deeper understanding of the real-world tradeoffs that arise when targeted differential privacy (and privacy protections more broadly) is used to make targeting decisions. In so doing, this can facilitate new opportunities for data access and data sharing, and help distill the complexities of data privacy into a collection of legible and measurable tradeoffs.

Methods

This section provides a brief description of the mathematical foundations of targeted differential privacy, our private projection algorithm, and the experimental methods and results. Full details of our approach, including formal definitions and proofs of the theorems referenced, can be found in the *Supplemental Information*.

Targeted Differential Privacy

Preliminaries. Differential privacy was first introduced in 2006 to enable the computation of privacy preserving statistics [20, 24]. At the root of differential privacy is a normative distinction about what constitutes a privacy violation in statistical learning [56]: inferring information about a population is not considered a privacy violation, provided the statistical analysis does not reveal "too much" about any individual.

Differential privacy achieves this through the concept of neighboring datasets. Let $||\cdot||_2$ denote the L_2-norm, and $\mathbb{L}_2^{n\times d}$ denote the set of of real $n \times d$ matrices where the L_2 norm of each row is at most 1. We say two databases $X, X' \in \mathbb{L}_2^{n\times d}$ are *classic neighbors* if they agree on exactly $n-1$ rows. Conceptually, a randomized algorithm is differentially private (DP) if the probability of creating a privatized dataset X_{priv} is essentially the same (up to ϵ and δ) for every pair of neighboring datasets.

Definition 1. A randomized algorithm A from $\mathbb{L}_2^{n\times d}$ to $\mathbb{O}$ satisfies (ϵ, δ)-*differential privacy* if for all classic neighbors $X, X' \in \mathbb{L}_2^{n\times d}$ and for all events $E \in \mathbb{O}$, $\mathbb{P}(A(X) \in E) \leq e^{\epsilon}\mathbb{P}(A(X') \in E) + \delta$.

Targeting applications are different from population-level analyses since the ability to target individuals involves determining their eligibility status. This suggests that we must adapt the notion of neighboring datasets to preserve sufficient information in X_{priv} to differentiate between different eligibility statuses. We say $X, X' \in \mathbb{L}_2^{n\times d}$ are *B-neighbors* if they are classic neighbors, and for the single row i where they disagree, $||X_i - X_i'||_2 \leq B$. When $B = 2$, B-neighbors coincides with classic neighbors. This motivates the definition of targeted differential privacy (TDP) below, which quantifies protection over B-neighboring datasets in place of classic neighboring datasets.

Definition 2. A randomized algorithm A from $\mathbb{L}_2^{n\times d}$ to $\mathbb{O}$ satisfies (B, ϵ, δ)-*targeted differential privacy* if for all B-neighbors $X, X' \in \mathbb{L}_2^{n\times d}$ and for all events $E \in \mathbb{O}$, $\mathbb{P}(A(X) \in E) \leq e^{\epsilon}\mathbb{P}(A(X') \in E) + \delta$. By definition, when $B = 2$ we recover (ϵ, δ)-differential privacy.

Through this alteration, the auxiliary parameter B ensures X_{priv} contains sufficient information to distinguish different individuals (i.e., those whose data are more than B apart), but insufficient information to learn the specific data values of similar individuals (those whose data are at most B apart).

Related contextual adaptations. There is a rich scholarly history of altering the definition of differential privacy to better align with context-specific goals outside of statistical domains. For example, the concept joint differential privacy has been applied matching and allocation problems to cope with the fact that these problems are provable impossible to solve under differential privacy [57, 58]. In another example, a variant of differential privacy was introduced to enable the aggregation of privacy preferences in a self-referential voting system [26]. And, the concept of geo-indistinguishability was introduced to adapt differential privacy to location-based systems [59]. In all of these cases, the concept of differential privacy was altered in a principled manner to be congruous with the context-specific goals of the task at hand.

The definition of targeted differential privacy follows in this tradition by using the notion of B-neighbors to better account for the context-specific goals of targeting. This notion of B-neighbors has been used previously in several settings unrelated to targeting [60]: in spatial data analysis, this requirement is called differential privacy under a neighborhood [61]; in privacy for compressed sensing applications, this definition is called constrained differential privacy [62]; and in privacy for moving-horizon estimators, this definition is referred to as adjacent differential privacy [60, 63]. The definition of targeted differential privacy is also similar in spirit to metric differential privacy [64], which is an alternative generalization of classic differential privacy for contexts where neighboring definitions are ill-suited (such as geolocation and smart-metering tasks). We differentiate from prior work by showing how the parameter B affects and enables targeting applications below.

Necessary conditions for accurate targeting. Using (B, ϵ, δ)-targeted differential privacy, we can formalize the tradeoffs that manifest in all targeting applications. To enable accurate targeting, we need a targeting process to behave "similarly" when X_{priv} is used in place of X. This is formalized by assigning sufficient probability γ to the event that an individual's eligibility status remains the unchanged when privatized data is used in place of original data. Theorem 1 in *Supplemental Information* proves that if this accuracy conditions holds, then B, ϵ, and δ must exhibit the following relationship:

$$\lceil 2B^{-1} \rceil \geq \lceil \epsilon^{-1} \ln(Q) \rceil$$

where

$$Q = \frac{\delta + \gamma(e^{\epsilon} - 1)}{\delta + (1 - \gamma)(e^{\epsilon} - 1)}$$

This theorem provides the necessary conditions that *every* targeted (and classic) differential privacy algorithm must obey in order to enable accurate targeting. This has two implications for the theory and practice of provably private targeting.

The first implication is that accurate targeting (as specified by γ) is provably impossible for classic differential privacy with small parameter values. To unpack this relationship, consider the case when $\delta = 0$. Then, $Q = \gamma/(1-\gamma)$, so the inequality above reduces to $\lceil 2B^{-1} \rceil \geq \lceil \epsilon^{-1} \ln(\gamma/(1-\gamma)) \rceil$. As we strengthen the privacy parameter ϵ to the point where $\epsilon < \ln(\gamma/(1-\gamma))$, $\lceil \epsilon^{-1} \ln(\gamma/(1-\gamma)) \rceil \geq 2$; hence, $\lceil 2B^{-1} \rceil \geq 2$, yielding $B < 2$. A similar result also holds when $\delta > 0$ (see Example 1 below). Thus, when ϵ and δ are small, classic differential privacy cannot permit accurate targeting.

Second, this result can be used proactively to provide a collection of potential candidate values for B, ϵ, and δ for *any* targeted differential privacy algorithm based on the accuracy needs of a situation. This follows by the contrapositive of Theorem 1: if (B, ϵ, δ) satisfy $\lceil 2B^{-1} \rceil < \lceil \epsilon^{-1} \ln(Q) \rceil$, then the accuracy bound does not hold. We present one such example below.

Example 1. A program official in Togo wants to use a TDP algorithm for their anti-poverty program with a value of B such that $2B^{-1} \in \mathbb{Z}$. To ensure humanitarian relief is correctly dispersed, they require 99% confidence that an individual's eligibility status will remain unchanged when X_{priv} is in place of the original data X. For such B and γ, the bound from Theorem 1 simplifies to

$$B \leq \frac{2}{\lceil \epsilon^{-1} \ln(Q) \rceil}$$

If the program official uses $\epsilon = 1$ and $\delta = 1 \times 10^{-4}$, then by they only need to consider $B \leq 0.4$, since no TDP algorithm exists to meet the accuracy bound with $B > 0.4$. Alternatively, if program officials are willing to use $\epsilon = 4$, then they only need to consider $B \leq 1$. △

In practice, the parameter B required to enable accurate targeting may need to be smaller than those produced by the bound in Theorem 1. Semantically speaking, this follows as Theorem 1 provides the *necessary* conditions for (B, ϵ, δ) to meet a desired accuracy bound, but not *sufficient* conditions. In general, the extent to which B must be reduced to invoke a sufficient condition for accurate targeting is task specific, and will depend on the underlying relationship between the original data's features and the targeting outcome in question, as well as the specifications of the targeting process and the specific TDP algorithm used.

Private Projection Algorithm

Our private projection algorithm (Algorithm 1) requires 7 inputs: the dataset X that we wish to privatize, 5 privacy parameters $(B, \epsilon_1, \epsilon_2, \delta_1, \delta_2)$, and a parameter k that represents the dimensionality of a mathematical projection used internally by our algorithm. The strategy of the algorithm is based on Gondara and Wong's random projection method [30], with three technical alternations (detailed in *Supplemental Information*). Our algorithm begins in Steps 1-3 by randomly map datapoints in X from $\mathbb{R}^d$ to $\mathbb{R}^k$ in a manner that satisfy $(B, \epsilon_1, \delta_1)$-TDP (Proposition 2 in *Supplemental Information*). In Steps 4-6, the algorithms map these points back to $\mathbb{R}^d$ using the data's covariance structure in a manner that satisfy $(B, \epsilon_2, \delta_2)$-TDP (Lemma 7 in *Supplemental Information*). Due to this modular design, our private projection algorithm satisfies $(B, \epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-targeted differential privacy (Theorem 2 of *Supplemental Information*).

Algorithm 1: Private Projection Algorithm

```
Input:
  • Dataset $X$
  • Privacy parameters $B \in (0, 2], \epsilon_1 > 0, \epsilon_2 \in (0, 1), \delta_1 \in (0, 1), \delta_2 \in (0, 1)$
  • Projection parameter $k$
Output: Privatized matrix $X_{priv}$
1 Sample $R \in \mathbb{R}^{d \times k}$, where each entry is sampled uniformly at random from $\{-1, 0, 1\}$.
2 Compute the projection $P = k^{-1}XR$
3 Privatize the projection via $P_{priv} = P + G$ where $G \in \mathbb{R}^{n \times k}$, $G_{i,j} \sim N(0, \sigma^2)$, and
      $\sigma = \frac{B}{\sqrt{k}} \sqrt{d \ln((2/3)(e-1)+1) - k^{-1} \ln(\delta_1/2)} \frac{\sqrt{2(\ln(1/\delta_1) + \epsilon_1)}}{\epsilon_1}$
4 Compute the privatized covariance matrix of $C_{priv} = X^T X + G$ where $G_{i,j} \sim N(0, \sigma^2)$ and
      $\sigma = 2B\sqrt{2 \ln(1.25/\delta_2)}/\epsilon_2$
  for all $i \geq j$, and $G_{i,j} = G_{j,i}$ for all $i < j$.
5 Compute the right singular vectors of $C_{priv}$. Call them $V^T$.
6 Return $X_{priv} = P_{priv}(V^T R)^{\dagger} V^T$, where † represents the pseudo-inverse.
```

Constructing TDP (and DP) datasets. We experimentally evaluate the impact our private projection algorithm induces when X_{priv} is used for targeting in place of X. In both of our case studies, we normalize each row of X to ensure the L_2 norm is at most 1. Hence, when $B = 2$ our algorithm generates X_{priv} with classic differential privacy guarantees. For our simulations, we define the *parameter grid* of our experiments as all possible 6-tuples $(B, \epsilon_1, \delta_1, \epsilon_2, \delta_2, k)$, where

- $B \in \{0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1, 2\}$
- $\epsilon_1 \in \{2, 3\}$
- $\delta_1 = \frac{2}{3}\delta$
- $\epsilon_2 \in \{0.5, 0.9999\}$
- $\delta_2 = \frac{1}{3}\delta$
- $k = 10^4$

and the specific value for δ used to compute δ_1 and δ_2 is determined by the number of individuals in each of the case studies. As a technical note, we use 0.9999 instead of 1 for the value of ϵ_2, as the Gaussian perturbation of Step 4 requires ϵ_2 in (0,1) [21, 25, 65] (see Lemma 7 of *Supplemental Information*).

Following best practices of statistical agencies that utilize differential privacy, we set δ to be less than the inverse of the number of individuals whose data is being privatized by an algorithm [66, 67]. In the Togolese anti-poverty experiments, we privatize the data all at once with our algorithm, and hence set $\delta = (n+1)^{-1}$, where n is the number of individuals in the Togolese dataset. In the Nigerian micro-lending experiments, we construct X_{priv} by first randomly partitioning the original dataset X into 6 sub-matrices of nearly equal size; we then apply the private projection algorithm to each sub-matrix, and combine the privatized sub-matrices together to form X_{priv}. Since our private projection algorithm is applied to disjoint datasets of size at most $\lceil n/6 \rceil$ (where n is the number of individuals in the Nigerian dataset), we set δ_1 and δ_2 using the value of $\delta = (\lceil n/6 \rceil + 1)^{-1}$ when we privatize each sub-matrix. By Lemma 3 in *Supplemental Information*, this parallelization strategy does not increase the privacy loss of our algorithm.

This partitioning approach improves both the runtime and accuracy of our algorithm. Parallelization removes computational barriers that arise when performing matrix operations on large datasets (especially as n, d, and k increase). Additionally, the parallel computation of privatized disjoint sub-matrices of X improves the accuracy of X_{priv} by decreasing the amount of noise introduced in Steps 3 and 4 of the algorithm (as δ_1 and δ_2 are inversely related to n).

k-anonymity baseline

As an additional point of comparison, we examine the program effectiveness in our two case studies when X_{priv} satisfies k-anonymity [22], another commonly applied privacy standard. We utilize the Mondrian algorithm [23] to construct a k-anonymous X_{priv}, which has been empirically shown by Slijepvcevic et al. (2021) to outperform other k-anonymous algorithms for machine learning tasks [68]. To preserve experimental consistency across the research literature, we use the same open-source Mondrian implementation as Slijepvcevic et al. (2021), which requires three inputs: the original dataset X, the privacy parameter k (not to be confused with the value of k in the private projection algorithm), and a list of *quasi-identifiers* in X (i.e., the names of the columns that need to be privatized). Since our private projection algorithm introduces noise to every column of X, we configure Mondrian to treat every column of X as a quasi-identifier. This enables a baseline level of consistency between the targeted (and classic) differential privacy experiments and k-anonymity experiments in the sense that every column of X is subject to privacy alterations. In all of our experiments, we consider values of $k \in \{2, ..., 10\}$.

Case Studies

Anti-poverty program in Togo

Following in Aiken et al. [6], we use a machine learning model to determine eligibility for Togo's Novissi (anti-poverty) cash transfer program based on mobile phone metadata. The algorithm is trained on a dataset that matches individual survey responses to data obtained from mobile phone operators in Togo. The survey makes it possible to observe the average daily consumption (per capita, purchasing-power-parity adjusted) of a nationally representative sample of roughly 4,200 mobile subscribers in Togo. We refer to this measure of consumption as our target variable y, which we will try to predict from data on how those subscribers use their mobile phones. The second dataset, X, obtained from Togo's two mobile phone operators, contains 10 "features" that quantify the phone use of each of the mobile phone subscribers in survey, such as the number of nighttime calls made, the entropy of the number of contacts they text during the weekday, and the percent of calls in specific prefectures. We normalize each column of this dataset of features X to have mean 0 and standard deviation 1, and normalize each row to be in the 10-dimensional L_2 ball. The target variable is normalized to be in the $[0, 1]$ interval. Additional statistical information on the features and targets can be found in Table S1 in *Supplemental Information*.

We use 5-fold cross-validation to train a ridge regression model to predict consumption y from mobile phone use X. To simulate the targeting of Novissi, we label an individual as eligible for the Novissi program if their predicted consumption is below the 29^{th} percentile of all predictions (this corresponds to the budget constraint of the actual program implemented by the government). For each fold, we record the average accuracy and average false positive rate. Using these estimates, we estimate the average number of "true poor" individuals (i.e., those whose actual consumption is below the 29^{th} percentile) who are incorrectly excluded from the program because their predicted consumption is above the 29^{th} percentile. To generate estimates of exclusion errors that are nationally representative, we scale the exclusion errors from the surveyed sample of 4,200 to Togo's estimated adult population (i.e., individuals at least 15 years of age) of approximately 4.95 million individuals. (According to the World Bank, Togo's estimated population in 2019 was 8,243,094 [69]; and based on estimates from the United Nations Population Fund, approximately 60% of individuals in Togo are at least 15 years of age [70]).

To show how targeted differential privacy affects the performance of this program, we simulate how targeting would have worked if X_{priv} had been used in place of X. Specifically, for each 6-tuple of the privacy parameters $(B, \epsilon_1, \delta_1, \epsilon_2, \delta_2, k)$ we perform 50 simulations whereby we generate X_{priv} using our private projection algorithm, and then use the same 5-fold cross-validation splits to compute the average accuracy and average false positive rate of the targeting procedure. From these estimates, we compute the average number of true poor excluded from the program at the national scale.

To simulate program performance under k-anonymity, we construct X_{priv} using the Mondrian algorithm for each privacy parameter $k \in \{2, ..., 10\}$. We then use the 5-fold cross validation process to compute the average accuracy and average false positive rate of the targeting procedure. Since Mondrian's algorithm does not utilize randomness, we only generate one X_{priv} for each value of k. Using these estimates, we compute the average number of true poor excluded from the program at the national scale, and report the results for the most accurate parameter value ($k = 2$).

Micro-lending platform in Nigeria

Our analysis of the privacy-program effectiveness tradeoff in the Nigerian micro-lending case study relies on three data sources. The first dataset X consists of 15 features of 20,788 individuals, which quantify how each individual uses their phone. The second dataset y characterizes individuals as low-risk or high-risk borrowers, as determined by an alternative credit score derived by the Nigerian lender. Our third dataset contains information on the size of the first loan an individual applied for, which we use to determine the profit or loss associated with repayment or default on the loan. All three datasets contain a hashed pseudonymous identifier that enables us to associate records across datasets. We normalize X and y in the same manner as the Togolese anti-poverty case study. Additional statistical information on the features and credit scores can be found in Table S1 in *Supplemental Information*.

We train a logistic regression model (using 5-fold cross validation) to predict an applicant's riskiness (and hence their eligibility for a loan) using the 15 features derived from their mobile phone metadata. To calculate the profit associated with lending decisions based on the privatized versus the original data, we assume that profit accrues to the lender when they offer loans to low-risk borrowers, and decreases when they offer loans to high-risk borrowers. Specifically, for a loan applicant j, let l_j denote the requested loan amount, r_j denote the lender's revenue on the loan, and i_j denote the interest on borrower j's loan. Denote the lender's profit from individual j as π_j, which is determined by the following four cases:

1. If the loan eligibility algorithm correctly classifies j as high risk, then the lender does not offer j a loan. In this case, the lender earns no profit, so $\pi_j = 0$.

2. If the loan eligibility algorithm correctly classifies j as a low risk, then the lender offers j a loan and j repays the entire amount. In this case, the lender's profit is the revenue generated by the loan, so $\pi_j = r_j$.

3. If the loan eligibility algorithm incorrectly classifies j as a low risk, then the lender offers j the loan but is unable to collect payment. For our analysis, we assume the worst-case outcome for the lender in which j repays none of the loan, and the lender loses the initial value of the loan, plus any interest they would have earned on they loan amount had they not extended the loan. Thus, $\pi_j = -(l_j + i_j)$.

4. If the loan eligibility algorithm incorrectly classifies j as a high risk, then the lender does not offer j a loan. In this setting, the lender does not receive the revenue they would have earned if they had offered the loan to the low risk applicant, and $\pi_j = -r_j$.

The program's profitability is given by $\pi = \sum_{j\in[n]} \pi_j$. Across all folds, we compute the program's average classification accuracy and the profitability.

To quantify the impact of using a privatized dataset X_{priv} in place of X, we follow the same experimental process used in the Togolese experiments with one notable alteration: to compute estimates for the targeted and classic differential privacy experiments, we use the partitioning strategy described in *Constructing TDP (and DP) datasets* to improve the runtime and accuracy of private projection algorithm.

Privacy Attacks

We measure the privacy afforded by our algorithm against three threats: singling-out attacks, attribute inference attacks, and distinguishing attacks. Each of these privacy measures — described in more detail below — quantifies an adversary's inability to infer information about the original dataset X. These measures range between 0 and 1, with higher values corresponding to higher levels of privacy protection. We do not examine the privacy protection afforded by the k-anonymous

Mondrian algorithm, as the targeting accuracy it exhibited was too stark to be of use in practice (see Figures 2(A) and (B)).

In all of our attacks, we assume the adversary has full access to the privatization algorithm, and can hence use any information about the algorithm as part of their attack strategy. This modeling choice is analogous to Kerckhoffs' Principle from cryptography, which states that "the cipher method must not be required to be secret, and it must be able to fall into the hands of the enemy without inconvenience" [71]. This is because cryptographic methods that depend on secrecy can be easily compromised if that secrecy is violated. As such, to illustrate the worse-case privacy scenario, we conduct our privacy attacks with full knowledge of the privatization algorithm to provide a realistic testing scenario.

Singling-out protection. Singling-out protection quantifies X_{priv}'s resilience to *isolation* (i.e., inferring the existence of a unique row in X containing a certain collection of values). As defined by Article 29 Working Party 216 of the GDPR, singling-out "corresponds to the possibility to isolate some or all records which identify an individual in the dataset" [36]. We follow the approach of *predicate singling-out* pioneered by Cohen and Nissim [40]: using both the privatized dataset X_{priv} and the privatization algorithm A, can an adversary infer a set of conditions that is satisfied by exactly 1 row in X? If so, then the adversary has the means to isolate a single row in X (and hence single out this individual). We define the singling-out protection of a dataset as the proportion of individuals we are unable to single-out.

We quantify the singling-out protection present in the original and privatized datasets in both of our case studies using the following methodology. In the original setting, the data X is unaltered, i.e., $X_{priv} = X$. Hence, an adversary can trivially single-out an individual by determining which rows in X_{priv} are unique. That is, for every row in X_{priv}, an adversary can construct the *identity conditions for row i* via "$x_i = (X_{priv})_i$ for all features i." Since $X_{priv} = X$, the identity conditions for row i singles-out a row in X if and only if $(X_{priv})_i$ is unique in X_{priv}. Therefore, the original dataset's singling-out protection score is given by the fraction of non-unique rows in X.

In contrast, when X_{priv} is given by our private projection algorithm, the presence of noise ensures that the identity conditions will almost surely fail to isolate a row. For this reason, we develop a new singling-out attack that leverages the design of our algorithm (called a *net attack*) that generate better predicates by modifying the identify conditions to account for the presence of noise in X_{priv}. Let the *d-dimensional net of size $\eta \in \mathbb{R}^d$*, denoted as $\text{Net}(\eta; X_{priv})$, be the set of $z \in \mathbb{R}^d$ such that each entry z_j is within distance η_j of any row in X_{priv}. Then a row X_i is singled-out by the net attack if and only if $X_i \in \text{Net}(\eta; X_{priv})$ and $|\text{Net}(\eta; X_{priv})| = 1$.

We generate candidate values of η based the standard deviation in each column of X_{priv}. The rationale for this approach stems from the following two observations: since the standard deviation measures the typical spread of datapoints in each column, the adversary could leverage some constant multiple of the standard deviation to generate possible net sizes; and since our algorithm introduces noise, the standard of each privatized column is likely larger than the standard of each original column.

We use five different constants in $\left\{\frac{1}{10}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\right\}$ to scale the column-wise standard deviations of X_{priv} to generate five different candidate values for η. For each parameter tuple in the parameter grid, we run our private projection algorithm 50 times and execute the net attack on each privatized dataset generated using each of the candidate values for η. For each η, we compute the average singling-out protection. To appropriately measure the privacy protection afforded to an individual, we report the protection afforded based on the adversary's most successful attack. Hence, we compute X_{priv}'s singling-out protection for a parameter tuple as the worst-case protection across all values of η.

Attribute inference protection. In contrast to singling-out protection, attribute inference protection quantifies X_{priv}'s resilience to feature reconstruction (i.e., inferring column values of X). As conceptualized in Article 29 Working Paper 216 of the GDPR, attribute inference refers to "the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes" [36]. Our framework of attribute inference protection follows the approach described in Giomi et al. [72], where an adversary has access to the privatized data X_{priv}, the privatization algorithm A, and h columns of the original data X on all n individuals; using these three objects, the adversary's goal is to infer the missing $d - h$ columns of X.

For our analysis, we consider $h \in \{1, \lceil d/2 \rceil, d-1\}$, where the columns available to the adversary are randomly selected when $h \neq d - 1$. For our attack method, we use the inference attack module in Anonymeter [72]. This attack method is similar to the approach pioneered by Narayanan and Shmatikov [73], and proceeds as follows. Using the h columns in the original dataset, the adversary performs a nearest neighbor search on the same h columns of the privatized data. Once this record is found in the privatized dataset, the adversary uses $d - h$ values in the privatized data as their guess for the $d - h$ values in the original data. We say the attack succeeds for individual i on unknown column j if the relative error between the estimate and the actual value is at most 5%, and define the protection score of a column as the proportion of individuals for which the attack fails. We follow the holdout-approach of Giomi et al. [72] and transform this protection score into a relative protection score: since X and X_{priv} share many statistical properties, inferences due to the preservation of statistical structure between their columns are qualitatively different than those that arise due to X_{priv}'s retention of precise information on individuals [74]; as such, this transformed score refines the original score by quantifying X_{priv}'s inference risk that cannot be explained by its statistical similarity to X (see *Supplemental Information* for technical details).

To determine the baseline level of protection in the original dataset, we use the following methodology. When $h \neq d - 1$, for every column of X we run Anonymeter 50 times, randomly selecting h other columns during each iteration, and compute the average relative protection score. When $h = d - 1$, for every column of X we run Anonymeter once (as there is no randomness involved in the selection of the other columns) and compute the relative protection score. To assess the protection of our privatized datasets, for each 6-tuple of algorithm parameters in the parameter grid, we replicate the above process 50 times (even when $h = d - 1$) to capture the sampling variability of Algorithm 1. We define the attribute inference protection of the original and privatized datasets to the lowest average protection score across all values of h and all columns j (i.e., protection against the most successful attack).

Distinguishing protection. Distinguishing protection is based on the mathematical construct of the *privacy loss random variable*, which quantifies an adversary's inability to infer which potential datasets likely produced X_{priv}. This protection is modeled by the following cryptographic game [43]. An adversary has access to X_{priv} and the algorithm A that generated it. Given any two datasets $\hat{X}$ and $\tilde{X}$ that differ in exactly one row, the adversary attempts to determine which dataset was more likely to create X_{priv}. The information-theoretic extent to which an adversary can do so is quantified by the privacy loss random variable. In the *Supplemental Information*, we construct our measure of distinguishing protection, which is proportional to the inverse of the mean of the privacy loss random variable. By definition, the distinguishing protection of the original dataset is 0. We derive the distinguishing protection formula for our private projection algorithm, and compute its value for every parameter tuple in our parameter grid.

Interpreting privacy protection scores. The first two measures of privacy protection are computed via privacy audits [75], where we attack datasets to determine their resilience against singling-out and attribute inference attacks. A methodological limitation of privacy audits stems from their empirical nature; they can only be used to surface examples of privacy risk, but cannot be used to show an approach is risk-free [56, 75]. For this reason, the privacy protection scores for singling-out and attribute inference protection should be interpreted as an upper bound on the actual privacy provided by X_{priv}. Despite this limitation, privacy audits can still be used to provide complementary measures of privacy protection for formal approaches (such as TDP and DP, whose protection is quantified via the parameters B, ϵ, and δ) to understand the privacy afforded against real-world attacks [75]. Distinguishing protection, in contrast, is computed via the mathematical properties of our algorithm, and provides a mathematical lower bound on the protection afforded.

Data availability

The mobile phone datasets from Togo and Nigeria contain detailed information on billions of mobile phone transactions. These data contain proprietary and confidential information belonging to a private telecommunications operator and cannot be publicly released. Upon reasonable request, we can provide information to accredited academic researchers about how to request the proprietary data from the telecommunications operator. Contact the corresponding author for any such requests.

References

[1] N. Jeffrie, "The Mobile Gender Gap Report 2023," *London: GSMA Intelligence*, 2023.

[2] M. Shanahan and K. Bahia, "The State of Mobile Internet Connectivity Report 2023," *GSMA, London, UK*, 2023.

[3] D. Eaves, R. Pope, and B. McGuire, "Government as a platform: how policy makers should think about the foundations of digital public infrastructure," *Kennedy School Review*, vol. 19, pp. 126–131, 2019.

[4] T. Hong, "Explainer: What is digital public infrastructure?," *Bill and Melinda Gates Foundation*, 2023.

[5] GSMA, "The Mobile Economy 2024," *London: GSMA Intelligence*, 2024.

[6] E. Aiken, S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock, "Machine learning and phone data can improve targeting of humanitarian aid," *Nature*, vol. 603, no. 7903, pp. 864–870, 2022.

[7] J. Robinson, D. S. Park, and J. E. Blumenstock, "The impact of digital credit in developing economies: A review of recent evidence," *KDI School of Pub Policy & Management Paper*, no. 23-04, 2023.

[8] S. Ovide, "The Nightmare of Our Snooping Phones," *The New York Times*, 2021.

[9] M. Boorstein, M. Iati, and A. Shin, "Top U.S. Catholic Church official resigns after cellphone data used to track him on Grindr and to gay bars," *The Washington Post*, 2021.

[10] O. Dube, J. Blumenstock, and M. Callen, "Measuring religion from behavior: Climate shocks and religious adherence in afghanistan," tech. rep., National Bureau of Economic Research, 2022.

[11] S. A. Thompson and C. Warzel, "Twelve million phones, one dataset, zero privacy," *The New York Times*, vol. 19, p. 2019, 2019.

[12] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the national academy of sciences*, vol. 106, no. 36, pp. 15274–15278, 2009.

[13] L. Taylor and D. Broeders, "In the name of development: Power, profit and the datafication of the global south," *Geoforum*, vol. 64, pp. 229–237, 2015.

[14] L. Mann, "Left to other peoples' devices? a political economy perspective on the big data revolution in development," *Development and Change*, vol. 49, no. 1, pp. 3–36, 2018.

[15] L. Taylor, "The price of certainty: How the politics of pandemic data demand an ethics of care," *Big Data & Society*, vol. 7, no. 2, p. 2053951720942539, 2020.

[16] J. E. Blumenstock and N. Kohli, "Big Data Privacy in Emerging Market Fintech and Financial Services: A Research Agenda," *CEGA Working Paper Series*, 2023.

[17] K. Lindert, T. G. Karippacheril, I. R. Caillava, and K. N. Chávez, *Sourcebook on the foundations of social protection delivery systems*. World Bank Publications, 2020.

[18] M. Grosh and J. L. Baker, “Proxy means tests for targeting social programs,” *Living standards measurement study working paper*, vol. 118, pp. 1–49, 1995.

[19] D. Björkegren and D. Grissen, “Behavior revealed in mobile phone usage predicts credit repayment,” *The World Bank Economic Review*, vol. 34, no. 3, pp. 618–634, 2020.

[20] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284, Springer, 2006.

[21] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, “Our data, ourselves: Privacy via distributed noise generation,” in *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pp. 486–503, Springer, 2006.

[22] L. Sweeney, “k-anonymity: A model for protecting privacy,” *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[23] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in *22nd International conference on data engineering (ICDE’06)*, pp. 25–25, IEEE, 2006.

[24] C. Dwork, N. Kohli, and D. Mulligan, “Differential privacy in practice: Expose your epsilons!,” *Journal of Privacy and Confidentiality*, vol. 9, no. 2, 2019.

[25] B. Balle, G. Barthe, and M. Gaboardi, “Privacy amplification by subsampling: Tight analyses via couplings and divergences,” *Advances in neural information processing systems*, vol. 31, 2018.

[26] N. Kohli and P. Laskowski, “Epsilon voting: Mechanism design for parameter selection in differential privacy,” in *2018 IEEE Symposium on Privacy-Aware Computing (PAC)*, pp. 19–30, IEEE, 2018.

[27] R. McKenna, G. Miklau, and D. Sheldon, “Winning the NIST Contest: A scalable and general approach to differentially private synthetic data,” *arXiv preprint arXiv:2108.04978*, 2021.

[28] K. Kenthapadi, A. Korolova, I. Mironov, and N. Mishra, “Privacy via the johnson-lindenstrauss transform,” *arXiv preprint arXiv:1204.2606*, 2012.

[29] J. Blocki, A. Blum, A. Datta, and O. Sheffet, “The johnson-lindenstrauss transform itself preserves differential privacy,” in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pp. 410–419, IEEE, 2012.

[30] L. Gondara and K. Wang, “Differentially private small dataset release using random projections,” in *Conference on Uncertainty in Artificial Intelligence*, pp. 639–648, PMLR, 2020.

[31] L. Chen, “Johnson-lindenstrauss transformation and random projection,” *JL. pdf*, 2015.

[32] M. Nabil, “Random projection and its applications,” *arXiv preprint arXiv:1710.03163*, 2017.

[33] S. Petti and A. Flaxman, “Differential privacy in the 2020 us census: what will it do? quantifying the accuracy/privacy tradeoff,” *Gates open research*, vol. 3, 2019.

[34] D. K. Mulligan, C. Koopman, and N. Doty, “Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2083, p. 20160118, 2016.

[35] A. Kozyreva, P. Lorenz-Spreen, R. Hertwig, S. Lewandowsky, and S. M. Herzog, "Public attitudes towards algorithmic personalization and use of personal data online: Evidence from germany, great britain, and the united states," *Humanities and Social Sciences Communications*, vol. 8, no. 1, pp. 1–11, 2021.

[36] "Article 29 Data Protection Working Party 216 Opinion 05/2014 on Anonymisation Techniques."

[37] United Nations Human Rights Office of the High Commissioner, "International Covenant on Civil and Political Rights," Adopted 16 December 1996.

[38] Electronic Frontier Foundation, "Necessary & Proportionate: International Principles on the Application of Human Rights Law to Communications Surveillance, 2014."

[39] N. Oliver, E. Letouzé, H. Sterly, S. Delataille, M. De Nadai, B. Lepri, R. Lambiotte, R. Benjamins, C. Cattuto, V. Colizza, *et al.*, "Mobile phone data and covid-19: Missing an opportunity?," *arXiv preprint arXiv:2003.12347*, 2020.

[40] A. Cohen and K. Nissim, "Towards formalizing the GDPR's notion of singling out," *Proceedings of the National Academy of Sciences*, vol. 117, no. 15, pp. 8344–8352, 2020.

[41] J. Jia and N. Z. Gong, "{AttriGuard}: A practical defense against attribute inference attacks via adversarial machine learning," in *27th USENIX Security Symposium (USENIX Security 18)*, pp. 513–529, 2018.

[42] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *28th USENIX Security Symposium (USENIX Security 19)*, pp. 1895–1912, 2019.

[43] M. Nasr, S. Songi, A. Thakurta, N. Papernot, and N. Carlin, "Adversary instantiation: Lower bounds for differentially private machine learning," in *2021 IEEE Symposium on security and privacy (SP)*, pp. 866–882, IEEE, 2021.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[45] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International conference on machine learning*, pp. 1225–1234, PMLR, 2016.

[46] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: a survey and review," *arXiv preprint arXiv:1412.7584*, 2014.

[47] E. Francis, J. Blumenstock, and J. Robinson, "Digital credit: A snapshot of the current landscape and open research questions," *CEGA White Paper*, pp. 1739–76, 2017.

[48] R. Burgess and R. Pande, "Do rural banks matter? evidence from the indian social banking experiment," *American economic review*, vol. 95, no. 3, pp. 780–795, 2005.

[49] H. Nissenbaum, "Privacy as contextual integrity," *Wash. L. Rev.*, vol. 79, p. 119, 2004.

[50] N. Kohli and P. Laskowski, "Differential privacy for black-box statistical analyses," *Proceedings on Privacy Enhancing Technologies*, vol. 3, pp. 418–431, 2023.

[51] C. Bowman, A. Gesher, J. K. Grant, D. Slate, and E. Lerner, *The architecture of privacy: On engineering technologies that can deliver trustworthy safeguards*. " O'Reilly Media, Inc.", 2015.

[52] J. A. Kroll, N. Kohli, and P. Laskowski, "Privacy and policy in polystores: a data management research agenda," in *Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB 2019 Workshops, Poly and DMAH, Los Angeles, CA, USA, August 30, 2019, Revised Selected Papers 5*, pp. 68–81, Springer, 2019.

[53] Z. Mumtaz and P. Whiteford, "Machine learning based approach for sustainable social protection policies in developing societies," *Mobile Networks and Applications*, vol. 26, pp. 159–173, 2021.

[54] E. Aiken, S. Bellue, J. Blumenstock, D. Karlan, and C. R. Udry, "Estimating impact with surveys versus digital traces: Evidence from randomized cash transfers in togo," tech. rep., National Bureau of Economic Research, 2023.

[55] K. Kirkpatrick, "Using algorithms to deliver disaster aid," *Communications of the ACM*, vol. 66, no. 6, pp. 17–19, 2023.

[56] N. Kohli, "Leveraging Differential Privacy While Attending to Social and Political Commitments," *PhD diss. University of California, Berkeley*, 2021.

[57] M. Kearns, M. Pai, A. Roth, and J. Ullman, "Mechanism design in large games: Incentives and privacy," in *Proceedings of the 5th conference on Innovations in theoretical computer science*, pp. 403–410, 2014.

[58] J. Hsu, Z. Huang, A. Roth, T. Roughgarden, and Z. S. Wu, "Private matchings and allocations," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 21–30, 2014.

[59] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 901–914, 2013.

[60] D. Desfontaines and B. Pejó, "SOK: Differential Privacies," *arXiv preprint arXiv:1906.01337*, 2019.

[61] C. Fang and E.-C. Chang, "Differential privacy with δ-neighbourhood for spatial and dynamic datasets," in *Proceedings of the 9th ACM symposium on Information, computer and communications security*, pp. 159–170, 2014.

[62] S. Zhou, K. Ligett, and L. Wasserman, "Differential privacy with compression," in *2009 IEEE International Symposium on Information Theory*, pp. 2718–2722, IEEE, 2009.

[63] V. Krishnan and S. Martínez, "A probabilistic framework for moving-horizon estimation: Stability and privacy guarantees," *IEEE Transactions on Automatic Control*, vol. 66, no. 4, pp. 1817–1824, 2020.

[64] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *Privacy Enhancing Technologies: 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings 13*, pp. 82–102, Springer, 2013.

[65] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, "Analyze gauss: optimal bounds for privacy-preserving principal component analysis," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 11–20, 2014.

[66] H. Page, C. Cabot, and K. Nissim, "Differential privacy an introduction for statistical agencies," *NSQR. Government Statistical Service*, pp. 1–53, 2018.

[67] J. M. Abowd, R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, *et al.*, "The 2020 census disclosure avoidance system topdown algorithm," *Harvard Data Science Review*, no. Special Issue 2, 2022.

[68] D. Slijepčević, M. Henzl, L. D. Klausner, T. Dam, P. Kieseberg, and M. Zeppelzauer, "k-anonymity in practice: How generalisation and suppression affect machine learning classifiers," *Computers & Security*, vol. 111, p. 102488, 2021.

[69] World Bank, "Population, Total - Togo." https://data.worldbank.org/indicator/SP.POP.TOTL?locations=TG.

[70] United Nations Population Fund, "World Population Dashboard Togo." https://www.unfpa.org/data/world-population/TG.

[71] J. Katz and Y. Lindell, *Introduction to modern cryptography: principles and protocols*. Chapman and hall/CRC, 2007.

[72] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, "A unified framework for quantifying privacy risk in synthetic data," *arXiv preprint arXiv:2211.10459*, 2022.

[73] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, IEEE, 2008.

[74] M. Bun, D. Desfontaines, C. Dwork, M. Naor, K. Nissim, A. Roth, A. Smith, T. Steinke, J. Ullman, and S. Vadhan, "Statistical inference is not a privacy violation." DifferentialPrivacy.org, 06 2021. https://differentialprivacy.org/inference-is-not-a-privacy-violation/.

[75] R. Cummings, D. Desfontaines, D. Evans, R. Geambasu, Y. Huang, M. Jagielski, P. Kairouz, G. Kamath, S. Oh, O. Ohrimenko, *et al.*, "Advancing differential privacy: Where we are now and future directions for real-world deployment," *Harvard Data Science Review*, 2024.

[76] A. Deaton, *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications, 1997.

[77] B. Bradbury, "Targeting social assistance," *Fiscal Studies*, vol. 25, no. 3, pp. 305–324, 2004.

[78] J. E. Blumenstock, "Fighting poverty with data," *Science*, vol. 353, no. 6301, pp. 753–754, 2016.

[79] E. L. Aiken, G. Bedoya, J. E. Blumenstock, and A. Coville, "Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in afghanistan," *Journal of Development Economics*, vol. 161, p. 103016, 2023.

[80] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[81] F. 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*, pp. 19–30, 2009.

[82] D. Achlioptas, "Database-friendly random projections," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001.

[83] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 245–250, 2001.

[84] D. Desfontaines, "The privacy loss random variable." https://desfontain.es/blog/privacy-loss-random-variable.html, 03 2020. Ted is writing things (personal blog).

Acknowledgements

This work was supported by the Lab for Inclusive FinTech (LIFT) at the University of California, Berkeley; the Bill and Melinda Gates Foundation via the Center for Effective Global Action; and the National Science Foundation under CAREER Grant IIS-1942702. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the U.S. Government. We are grateful to Emily Aiken, Suraj Nair, Dan Cassara, Paul Laskowski, and Seth Garz for providing early feedback on this work.

Author contributions statement

JB conceived of the research idea. NK developed the algorithm, proved the theorems, and conducted the empirical measurements. JB and NK wrote the manuscript.

Competing interests

The author(s) declare no competing interests.

Materials and correspondence

Correspondence and requests for materials should be addressed to Joshua E. Blumenstock.

Figures

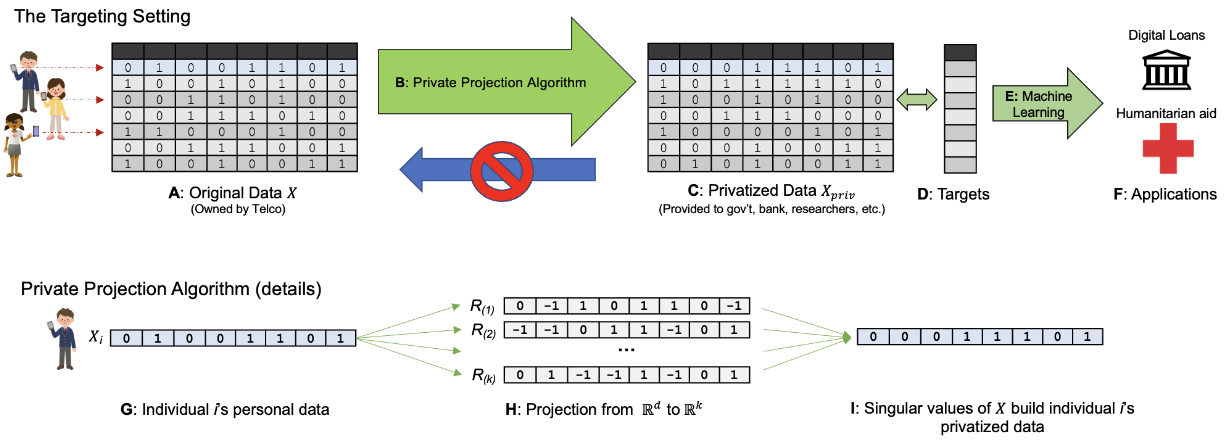

Figure 1: Overview of targeting setting (top) and the private projection algorithm (bottom). (A) Personal data are held by a data holder (e.g., mobile network operator). These data are given to (B) an algorithm that generates a version of the data with a provable privacy guarantee. (C) This private version of the data can then be joined with (D) training "labels" from third parties that indicate the true eligibility for a subset of the population. (E) Machine learning models learn how to predict eligibility status for the full population for whom eligibility status is not directly observed, but for whom private data exist. (F) These predictions can then be used in downstream applications. Bottom figures provide a conceptual sketch of how the private projection algorithm works. (G) Each individual's raw data X is projected into $\mathbb{R}^k$ (H), and is then projected back to $\mathbb{R}^d$ using the singular values of X (I). The resulting record corresponds to individual i's record in X_{priv}.

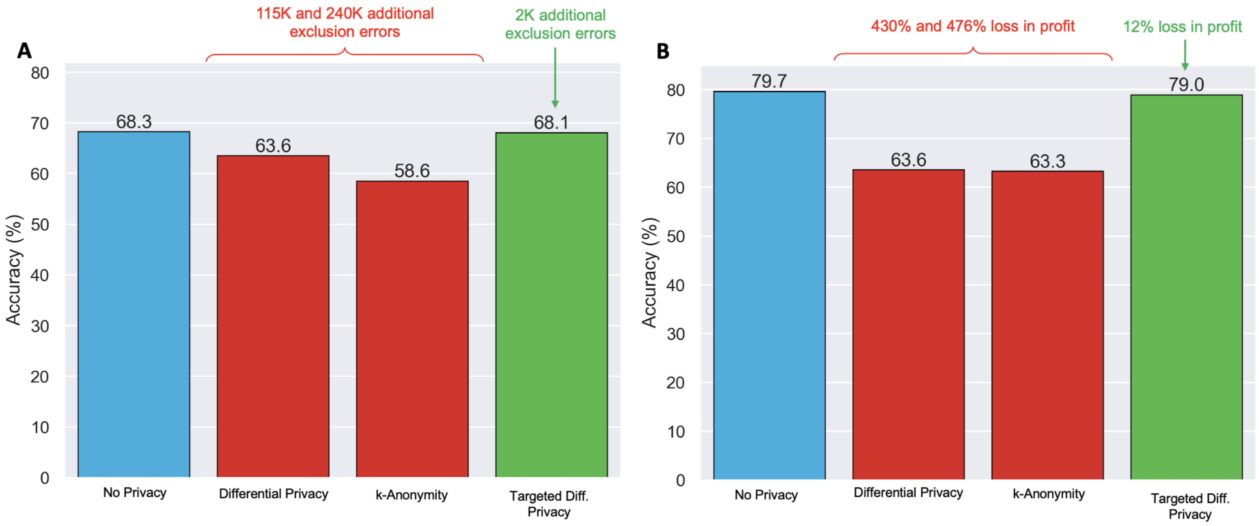

Figure 2: The impact of different privacy-enhancing technologies on targeting accuracy in two real-world settings. (A) Anti-poverty program in Togo. In simulations of a nationwide humanitarian program, differential privacy ($\epsilon \approx 4$) and k-anonymity ($k = 2$) would increase exclusion errors by $115K$ and $240K$, relative to the non-private status quo. Our approach (targeted differential privacy) increases exclusion errors by $2K$ ($B = 0.25$). (B) Micro-lending platform in Nigeria. Simulating the accuracy of credit scoring algorithms used to extend loans to individuals without a formal financial history, differential privacy ($\epsilon \approx 3$) and k-anonymity ($k = 2$) would reduce the profits of the program by 430% and 476%, respectively. Targeted differential privacy would reduce profits by 12% ($B = 0.1$).

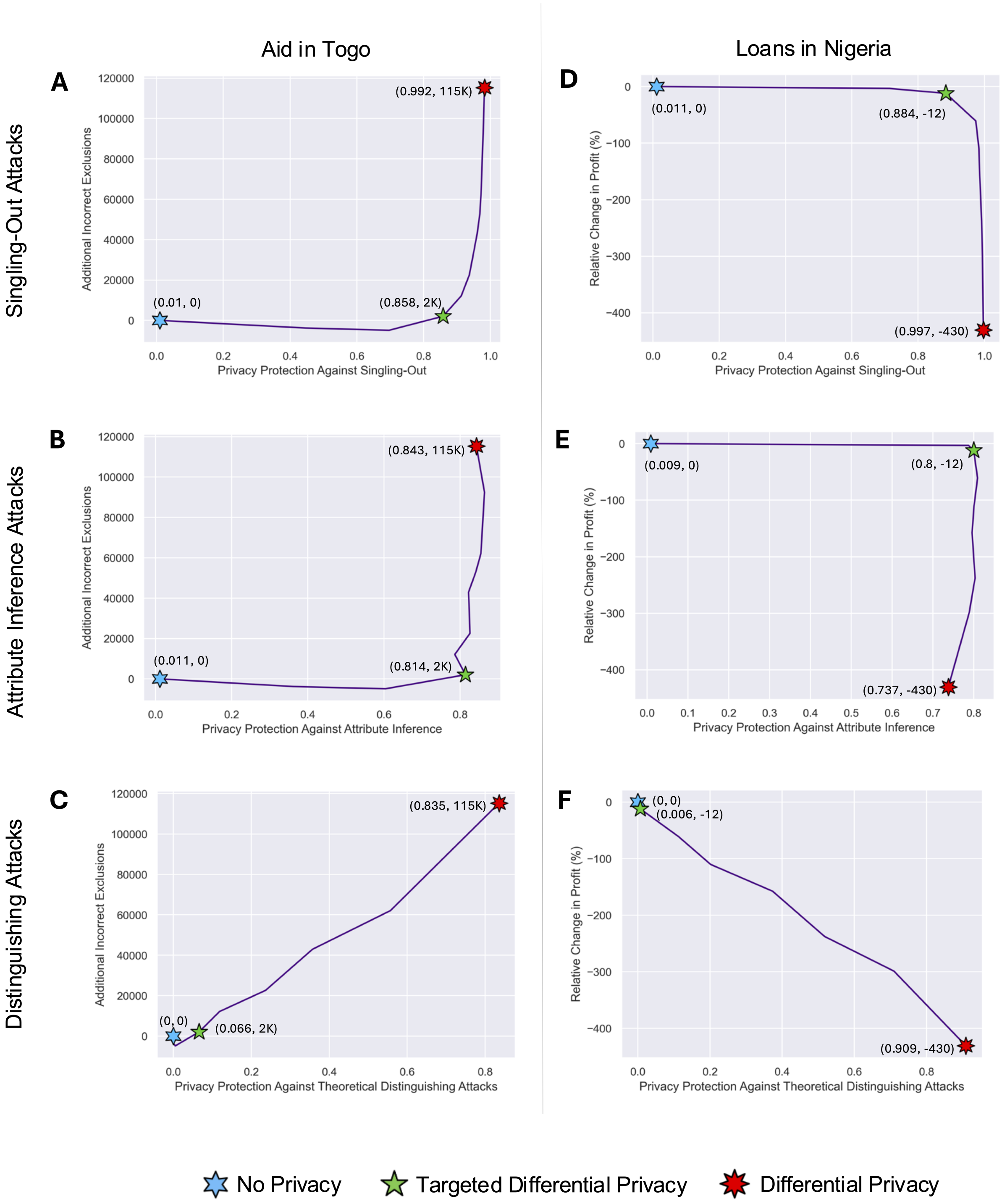

Figure 3: Empirical tradeoffs between privacy and program effectiveness for a humanitarian program in Togo (left panels) and a digital lending platform in Nigeria (right panels). Privacy protections are shown for singling-out attacks (top row), attribute inference attacks (middle row), and distinguishing attacks (bottom row). For all panels, blue stars represents the non-private status quo, and red stars represents classic (ϵ, δ)-differential privacy. Green stars corresponds to targeted differential privacy, with $B = 0.25$ and $\epsilon \approx 4$ for the Togolese program and $B = 0.1$ and $\epsilon \approx 3$ for the Nigerian application.

Supplemental Information

1 Supplementary Note: Targeted Differential Privacy

1.1 Formalizing the targeting problem and setting

Our framework of targeted differential privacy is motivated by an interaction between a data holder (such as a mobile network operator) and a downstream program official who wishes to make use of the data. The program official seeks to target a subset of n individuals for some benefit, based on a *targeting variable* y. This targeting variable could be an individual's or household's income [76, 77], their daily or household consumption [78, 6], or their credit score [19]. However, the program official faces an information restriction: they only possess target variables for a non-empty proper subset $I \subset [n] = \{1, ..., n\}$ (these individuals are *in the training sample*). Call these targets $y_I \in \mathbb{R}^{|I|}$. Let $O = [n] \cap I^c$ denote the subset of individuals *out of the training sample*.

In certain humanitarian situations, it is infeasible to directly solicit targeting information on O [78, 6, 79]. To overcome these informational constraints, program officials have partnered with data holders (such as a mobile network operator) to solicit data on individuals in $[n]$ to predict these targeting variables using techniques from machine learning [78, 6, 79]. We represent such data as an $n \times d$ matrix X, where each row corresponds to d features about individual $i \in [n]$. For notational purposes, we let X_i denote the i^{th} row of X, and $X_{(j)}$ denote the j^{th} column of X. Following this convention, we let X_I denote the row-wise submatrix of X that contains data on the individuals in I. For ease of exposition, we will use the terms matrices, datasets, and databases interchangeably.

Equipped with X_I with y_I, the program official constructs a machine learning model M, which can then be used to generate *predicted targeting variables* $\hat{y}$. In practice, a value of $\hat{y}$ is generated for every individual in $[n]$, even if their actual targeting variable is known to the program official. The rationale is that this provides a level of fairness in the sense that the eligibility of all individuals is determined by the same targeting process.

For our study, we consider the setting where the data holder – cognizant of privacy concerns – does not share a dataset X with the program official.[1] Instead, the data holder provides the program official with a *provably private* dataset X_{priv}, which will be used in place of X in the learning and prediction process described above. The dataset X_{priv}, which we will refer to as a *privatized dataset*, is the output of some privacy-enhancing technology A.

Our study seeks to understand the properties A must have to ensure X_{priv} (1) yields accurate predictions to enable effective programmatic outcomes, while (2) providing strong privacy guarantees, as well as characterize the limits and tradeoffs induced by A.

1.2 Targeted differential privacy

To describe targeted differential privacy, we will utilize the following notation and terminology. Let $||\cdot||_2$ denote the L_2-norm, $\mathbb{L}_2^d$ denote the the set of d-dimensional real vectors with L_2 at most 1, and $\mathbb{L}_2^{n\times d}$ denote the set of of real $n \times d$ matrices where the L_2 norm of each row is at most 1.

We say two databases $X, X' \in \mathbb{L}_2^{n\times d}$ are *classic neighbors* if they agree on exactly $n-1$ rows. Classic differential privacy ensures that an algorithm's behavior is statistically indistinguishable

[1] In the setting we consider, the program official does not share y_I with the data holder. In situations where the data holder can share y_I as part of a secure computing protocol, it is possible to utilize multiple privacy-enhancing technologies to generate $\hat{y}$. For example, one possible design solution would be to create a differentially private machine learning model M using secure multiparty computation, and then predict each individual's targeting variable using X.

for all pairs of classically neighboring datasets [21].

Definition 3. A randomized algorithm A from $\mathbb{L}_2^{n\times d}$ to $\mathbb{O}$ satisfies (ϵ, δ)-*classic differential privacy* if for all classic neighbors $X, X' \in \mathbb{L}_2^{n\times d}$ and for all measurable sets $E \in \mathbb{O}$,

$$\mathbb{P}(A(X) \in E) \leq e^{\epsilon}\mathbb{P}(A(X') \in E) + \delta$$

For brevity, at times we will omit the term "classic" and simply refer to this condition as *differential privacy* (DP).

For machine learning applications, the specific values in the row of a dataset can affect the predictions made. As such, it will be helpful to consider a more granular notion of neighboring datasets that quantities the extent to which two rows disagree.

We say $X, X' \in \mathbb{L}_2^{n\times d}$ are *B-neighbors* if they are classic neighbors, and for the single row i where they disagree, $||X_i - X_i'||_2 \leq B$. When $B = 2$, B-neighbors coincides with classic neighbors. As we will see, B serves as an analytical device, enabling a finer analysis of privacy loss compared to classic differential privacy.

We define the cumulative distance between databases X and X', both of size n, as the sum of their row-wise distances. That is, if X and X' have n rows, their cumulative distance is $d(X, X') = \sum_{i\in[n]} ||X_i - X_i'||_2$. By definition, if X and X' are B-neighbors, then $d(X, X') \leq B$.

Targeted differential privacy is a variant of classic differential privacy that requires an algorithm's behavior satisfy statistical indistinguishability for all pairs of B-neighboring datasets.

Definition 4. A randomized algorithm A from $\mathbb{L}_2^{n\times d}$ to $\mathbb{O}$ satisfies (B, ϵ, δ)-*targeted differential privacy*[2] (TDP) if for all B-neighbors $X, X' \in \mathbb{L}_2^{n\times d}$ and for all measurable sets $E \in \mathbb{O}$,

$$\mathbb{P}(A(X) \in E) \leq e^{\epsilon}\mathbb{P}(A(X') \in E) + \delta$$

By definition, when $B = 2$ we recover classic differential privacy.

The proofs of post-processing, sequential composition, and parallel composition are nearly identical to those presented in Dwork and Roth [80] and McSherry [81], so we omit them.

Lemma 1. *[Post-Processing] Suppose A from $\mathbb{L}_2^{n\times d}$ to $\mathbb{O}$ satisfies (B, ϵ, δ)-TDP. Then for any (potentially randomized) function F from $\mathbb{O}$ to some space $\mathbb{O}'$, $F \circ A$ also satisfies (B, ϵ, δ)-TDP.*

Lemma 2. *[Sequential Composition] Suppose randomized algorithms A_j mapping $\mathbb{L}_2^{n\times d}$ to $\mathbb{O}_j$ satisfies $(B_j, \epsilon_j, \delta_j)$-TDP for $j \in [m]$. Then the sequential composition of these mechanisms, $(A_1, ..., A_m)$ from $\mathbb{L}_2^{n\times d}$ to $\prod_{j\in[m]} O_j$ satisfies $(\min_{j\in[m]} B_j, \sum_{j\in[m]} \epsilon_j, \sum_{j\in[m]} \delta_j) - TDP$*

Lemma 3. *[Parallel Composition] Suppose A_j from $\mathbb{L}_2^{n\times d}$ to $\mathbb{O}_j$ satisfies $(B_j, \epsilon_j, \delta_j)$-TDP. If these algorithms are computed on disjoint sets of data, the composite mechanism $(A_1, ..., A_m)$ from $\mathbb{L}_2^{n\times d}$ to $\prod_{j\in[m]} \mathbb{O}_j$ satisfies $(\min_{j\in[m]} B_j, \max_{j\in[m]} \epsilon_j, \max_{j\in[m]} \delta_j)$-TDP.*

Privacy protection under TDP decays predictably for datasets x and y that are classic neighbors.

Lemma 4. *[Switch Lemma] Suppose a randomized algorithm A from $\mathbb{L}_2^{n\times d}$ to $\mathbb{O}$ satisfies (B, ϵ, δ)-TDP with $B, \epsilon > 0$. Then for all classic neighboring datasets $X, \hat{X} \in \mathbb{L}_2^{n\times d}$, and for all measurable sets $E \in \mathbb{O}$,*

$$\mathbb{P}(A(X) \in E) \leq e^{s\epsilon}\mathbb{P}(A(\hat{X}) \in E) + \frac{e^{s\epsilon} - 1}{e^{\epsilon} - 1}\delta$$

where $s = \lceil d(X, \hat{X})B^{-1} \rceil$

[2] The definition of TDP has appeared in prior privacy studies in different contexts. For additional information, see *Related Contextual Adaptations* in the *Methods* section of the main manuscript.

Proof. Consider an arbitrary measurable set $E \subseteq \mathbb{O}$. If $d(X, \hat{X}) \leq B$, the claim follows immediately. So consider $d(X, \hat{X}) > B$. Since X and $\hat{X}$ are classic neighbors, they differ in one element: denote this as the the t^{th} element. Then $||X_t - \hat{X}_t||_2 = d(X, \hat{X})$.

Next, we will relate X and $\hat{X}$ to an sequence of B-neighboring datasets; to do so, we will first construct a preliminary sequence of datapoints starting at X_t and ending at $\hat{X}_t$ whose adjacent values in the sequence are at most distance B apart. To motivate the construction used in the upcoming paragraph, consider the line segment in $\mathbb{L}_2^d$ whose endpoints are X_t and $\hat{X}_t$. This line segment has length $d(X, \hat{X})$. We will break this line segment into $m + 1$ segments that are each at most length B. Since $\mathbb{L}_2^d$ is convex, these breakpoints must reside in $\mathbb{L}_2^d$, and will be used as values in our preliminary sequence.

This is done mathematically by letting m be the smallest non-negative integer such that $d(X, \hat{X}) - mB \leq B$. Then m is the smallest non-negative integer such that $m \geq d(X, \hat{X})B^{-1} - 1$. So $m = \lceil d(X, \hat{X})B^{-1} - 1 \rceil$, which implies $m + 1 = \lceil d(X, \hat{X})B^{-1} \rceil$. Consider a sequence $X_t^{(0)}, X_t^{(1)}, ..., X_t^{(m+1)} \in \mathbb{L}_2^d$, defined as follows: $X_t^{(0)} = X_t$, $X_t^{(m+1)} = \hat{X}_t$, and the remaining $X_t^{(i)}$ are any values such that $||X_t^{(i)} - X_t^{(i+1)}||_2 \leq B$ for $i \in \{0, ..., m\}$, provided they are consistent with the endpoints $X_t^{(0)}$ and $X_t^{(m+1)}$.

Now, define $X^{(i)}$ as the dataset X with X_t replaced with $X_t^{(i)}$. This constructs the sequence of datasets $X^{(0)}, X^{(1)}, ..., X^{(m+1)}$. By construction, all of these datasets are in $\mathbb{L}_2^{n \times d}$; also, $X = X^{(0)}$ and $\hat{X} = X^{(m+1)}$.

Write $p_i = \mathbb{P}(A(X^{(i)}) \in E)$. By construction, every adjacent pair of datasets $X^{(i)}$ and $X^{(i+1)}$ in our sequence are B-neighbors, so

$$\begin{aligned} p_0 &\leq e^{\epsilon} p_1 + \delta \\ &\leq e^{\epsilon}(e^{\epsilon} p_2 + \delta) + \delta \\ &= e^{2\epsilon} p_2 + (e^{\epsilon} + 1)\delta \\ &\dots \\ &\leq e^{(m+1)\epsilon} p_{m+1} + (e^{m\epsilon} + \dots + e^{\epsilon} + 1)\delta \\ &= e^{(m+1)\epsilon} p_{m+1} + \frac{e^{(m+1)\epsilon} - 1}{e^{\epsilon} - 1}\delta \end{aligned}$$

Since $m + 1 = s$, $p_0 = \mathbb{P}(A(X) \in E)$, and $p_{m+1} = \mathbb{P}(A(\hat{X}) \in E)$, the claim is proven. □

The switch lemma provides a bridge between targeted and classical differential privacy. Namely, if A satisfies (B, ϵ, δ)-TDP then A also satisfies (ϵ', δ')-DP, where $\epsilon' = s\epsilon$, $\delta' = \min\left\{1, \frac{e^{s\epsilon}-1}{e^{\epsilon}-1}\delta\right\}$, and $s = \lceil 2B^{-1} \rceil$. However, the converse is not necessarily true: TDP requires the (ϵ, δ) probability bound holds for every diameter B ball; however, (ϵ', δ')-DP does not necessarily imply the (ϵ, δ) probability bound holds in every diameter B ball. As such, there is a key semantic distinction between classic and targeted differential privacy: while TDP guarantees an adversary cannot discern (up to the factors ϵ and δ) between two individuals' datapoints that are B-similar, under (ϵ', δ')-DP an adversary may be able to do so.

1.3 Formal analysis of the necessary conditions for accurate targeting

In this section, we formalize the relationship between the privacy parameters (B, ϵ, δ) and the the goal of accurate targeting. We consider the binary setting, where an individual is classified either as eligible for a benefit ($y = 1$) or not ($y = 0$). Denote the set $\{0, 1\}$ as $\mathbb{Z}_2$.

Let $T : \mathbb{L}_2^{n\times d} \longrightarrow \mathbb{Z}_2$ denote the targeting process for an arbitrary individual. In the setting we consider in the main manuscript, $T(X)$ is the result of running a learning algorithm on (X_I, y_I), outputting the machine learning model, and then using this model to determine an individual's eligibility status. When $A(X)$ outputs the privatized dataset X_{priv} that is used in place of X for targeting, the resulting eligibility classification is given by $T(A(X))$.

We restrict our analysis to targeting processes T that are *minimally responsive*: these are functions for which there exist classic neighboring datasets X and X' such that $T(X) \neq T(X')$. Note that if T is not minimally responsive, then T outputs the same predictions regardless of an individual's data. Such targeting processes are not useful in practice, and hence we omit them from our analysis.

To enable accurate targeting, we need the mechanism $T \circ A$ to behave similarly to T with high probability. Formally, we say A is *γ-accurate* for T if for all $X \in \mathbb{L}_2^{n\times d}$, we have $\mathbb{P}(T(A(X)) = T(X)) \geq \gamma$.

We now characterize the necessary conditions for accurate targeting under minimally responsive targeting processes.

Theorem 1. *Let A be a randomized algorithm from $\mathbb{L}_2^{n\times d}$ to $\mathbb{L}_2^{n\times d}$ satisfying (B, ϵ, δ)-TDP with $B, \epsilon > 0$. Suppose that $T : \mathbb{L}_2^{n\times d} \longrightarrow \mathbb{Z}_2$ is a minimally responsive deterministic function. If A is γ-accurate for T with $\gamma \in \left[\frac{1}{2}, 1\right)$, then*

$$\lceil 2B^{-1} \rceil \geq \lceil \epsilon^{-1} \ln(Q) \rceil$$

where

$$Q = \frac{\delta + \gamma(e^{\epsilon} - 1)}{\delta + (1-\gamma)(e^{\epsilon} - 1)}$$

In particular, when $2B^{-1} \in \mathbb{Z}$ and $\gamma > \frac{1}{2}$, the inequality simplifies to

$$B \leq \frac{2}{\lceil \epsilon^{-1} \ln(Q) \rceil}$$

Proof. By the minimal responsiveness of T, there exists classic neighboring datasets X and X' in $\mathbb{L}_2^{n\times d}$ such that $T(X) \neq T(X')$. Since A satisfies (B, ϵ, δ)-TDP, so too does $T \circ A$ by the post-processing lemma (Lemma 1). By the switch lemma (Lemma 4), we have

$$\mathbb{P}(T(A(X)) \in E) \leq e^{s\epsilon}\mathbb{P}(T(A(X')) \in E) + \frac{e^{s\epsilon} - 1}{e^{\epsilon} - 1}\delta$$

for all sets $E \subseteq \mathbb{Z}_2$, where $s = \lceil d(X, X')B^{-1} \rceil$. In particular, consider the set $E = \{T(X)\}$.

By the accuracy condition, $\mathbb{P}(T(A(X)) \in E) \geq \gamma$. Since $E^c = \{T(X')\}$, the accuracy condition also implies $\mathbb{P}(T(A(X')) \in E^c) \geq \gamma$, so $\mathbb{P}(T(A(X')) \in E) = 1 - \mathbb{P}(T(A(X')) \in E^c) \leq 1 - \gamma$. Combining this with the above inequality, we have

$$\gamma \leq e^{s\epsilon}(1-\gamma) + \frac{e^{s\epsilon} - 1}{e^{\epsilon} - 1}\delta$$

Rearranging terms and solving for s yields $s \geq \epsilon^{-1}\ln(Q)$. Note that the quantity on the right-hand side of the inequality is well-defined, as $\gamma \geq \frac{1}{2}$ implies $Q \geq 1$. Next, we deduce that $s \geq \lceil \epsilon^{-1} \ln(Q) \rceil$. There are two cases to consider.

- ⋆ *Case 1*: If $\epsilon^{-1}\ln(Q) \in \mathbb{Z}$, then $\epsilon^{-1}\ln(Q) = \lceil \epsilon^{-1}\ln(Q) \rceil$, so $s \geq \lceil \epsilon^{-1}\ln(Q) \rceil$ holds.
- ⋆ *Case 2*: Otherwise, if $\epsilon^{-1}\ln(Q) \in \mathbb{R} \cap \mathbb{Z}^c$, then $s \neq \epsilon^{-1}\ln(Q)$ since the left-hand side is an integer and the right hand side is not. Hence we must have $s > \epsilon^{-1}\ln(Q)$, implying $s \geq \lceil \epsilon^{-1}\ln(Q) \rceil$.

Therefore, in either case we have $s \geq \lceil \epsilon^{-1} \ln(Q) \rceil$. Since $s = \lceil d(X, X')B^{-1} \rceil$ and $d(X, X') \leq 2$, we have $s \leq \lceil 2B^{-1} \rceil$ because the ceiling function is monotonically non-decreasing. Therefore $\lceil 2B^{-1} \rceil \geq \lceil \epsilon^{-1} \ln(Q) \rceil$ as claimed.

Furthermore, when $2B^{-1} \in \mathbb{Z}$, the inequality simplifies to $2B^{-1} \geq \lceil \epsilon^{-1} \ln(Q) \rceil$. Also, $\gamma > \frac{1}{2}$ implies $\ln(Q) > 0$, so $\lceil \epsilon^{-1} \ln(Q) \rceil \geq 1$. Solving for B yields

$$B \leq \frac{2}{\lceil \epsilon^{-1} \ln(Q) \rceil}$$

□

1.4 Private projection algorithm details

Next, we describe an algorithm that satisfies (B, ϵ, δ)-TDP. Algorithm 1 is motivated by the Johnson-Lindenstrauss lemma [31, 32], and its differentially private variants [28, 29, 30]. Algorithm 1 is based on Gondara and Wong's differentially private randomized projection method [30], with three notable alterations (two of which adapt to the projection phase, and the remaining adapts the covariance projection stage). The first alteration is in Step 1 of the algorithm. Gondara and Wong sample the elements of R from a Gaussian with mean 0 and variance k^{-1}. Our algorithm instead samples values from $\{-1, 0, 1\}$ uniformly at random. This improves the practical efficiency of our algorithm by utilizing integer computations over floating point computations [32, 82, 83]. The second alteration occurs in Step 2 of the algorithm where the projection values are inversely scaled by k, which reduces the standard deviation of the noise used as k increases in Step 3 to achieve TDP. The third alteration occurs in Step 4 of the algorithm. In their algorithm, Gondara and Wong construct a noisy covariance matrix C_{priv} by summing together the covariance matrix $X^T X$ and a matrix whose G whose values are drawn from a Gaussian distribution. Since the matrix G may not be symmetric, their resulting C_{priv} may not be symmetric. For this reason we follow the approach in Dwork et al. [65] and construct a perturbation matrix G whose upper triangle values are sampled from a Gaussian distribution, and whose lower triangular values are copied from the upper triangle. By construction, G is now symmetric. Since the sum of symmetric matrices is symmetric, this design alteration has the benefit of constructing a symmetric C_{priv}.

The remainder of this section is devoted to showing that Algorithm 1 satisfies (B, ϵ, δ)-TDP. We begin with the following lemma, which examines the sensitivity of the matrix multiplication by R in Step 2, as measured by the Frobenius norm $|| \cdot ||_F$.

Lemma 5. *Suppose $X, X' \in \mathbb{L}_2^{n \times d}$ are B-neighbors. For any matrix $R \in \mathbb{R}^{d \times k}$,*

$$||k^{-1}XR - k^{-1}X'R||_F^2 \leq \frac{B^2}{k^2} \sum_{j=1}^{k} \sum_{p=1}^{d} R_{p,j}^2$$

Proof. Let $R \in \mathbb{R}^{d \times k}$ be given. Since X and X' are B-neighbors, there exists a unique row, say $i \in [n]$, such that $X_i - X'_i \neq 0$. Consider the (p, j) entry of $XR - X'R$. This is given by $\mathbb{1}(p = i)(X_p - X'_p)R_{(j)}$.

So then

$$
\begin{aligned}
||XR - X'R||_F^2 &= \sum_{j=1}^{k} |(X_i - X_i')R_{(j)}|^2 \\
&\leq \sum_{j=1}^{k} ||X_i - X_i'||_2^2 ||R_{(j)}||_2^2 \\
&= ||X_i - X_i'||_2^2 \sum_{j=1}^{k} ||R_{(j)}||_2^2 \\
&\leq B^2 \sum_{j=1}^{k} ||R_{(j)}||_2^2 \\
&= B^2 \sum_{j=1}^{k} \sum_{p=1}^{d} R_{p,j}^2
\end{aligned}
$$

where the first inequality follows by the Cauchy-Schwartz inequality and the second follows by the definition of B-neighbors. Hence, $||XR - X'R||_F^2 \leq B^2 \sum_{j=1}^{k} \sum_{p=1}^{d} R_{p,j}^2$. Multiplying both sides of the inequality by k^{-2} produces the claim. □

In particular, when $R \in \{-1, 0, 1\}^{d \times k}$, the inequality in Lemma 5 reduces to $||k^{-1}XR - k^{-1}X'R||_F \leq B\sqrt{d/k}$ in the worst-case. Rather than scale the noise in Step 3 of our algorithm using this worst-case bound, we can reduce the amount of noise introduced by considering a probabilistic version of the Frobenius norm based on the randomness induced by R. That is, we can instead guarantee $||k^{-1}XR - k^{-1}X'R||_F \leq \alpha$ except with some failure probability β, which we incorporate in our privacy analysis.

Proposition 1. *Suppose each element of $R \in \mathbb{R}^{d \times k}$ is sampled i.i.d from a categorical distribution over $\{-1, 0, 1\}$ as follows: for $p_0 \in (0, 1)$,*

z	-1	0	1
$\mathbb{P}(z)$	$\frac{1-p_0}{2}$	p_0	$\frac{1-p_0}{2}$

Then, $\mathbb{P}(||k^{-1}XR - k^{-1}X'R||_F \geq \alpha) \leq \beta$ whenever

$$
\alpha \geq \frac{B}{\sqrt{k}} \sqrt{d \ln((1 - p_0)(e - 1) + 1) - k^{-1} \ln(\beta)}
$$

Proof. Suppose the (p, j) entry of R is chosen according to the distribution described in the proposition. Then $R_{p,j}^2$ is 0 with probability p_0 and 1 with probability $1 - p_0$, so $R_{p,j}^2 \sim \text{Bernoulli}(1 - p_0)$. Let $Y = \sum_{j=1}^{k} \sum_{p=1}^{d} R_{p,j}^2$. Then $Y \sim \text{Binomial}(dk, 1 - p_0)$.

Next, observe that $\mathbb{P}(||k^{-1}XR - k^{-1}X'R||_F \geq \alpha) = \mathbb{P}(||k^{-1}XR - k^{-1}X'R||_F^2 \geq \alpha^2)$. By Lemma 5, if $||k^{-1}XR - k^{-1}X'R||_F^2 \geq \alpha^2$ then $B^2 k^{-2} \sum_{j=1}^{k} \sum_{p=1}^{d} R_{p,j}^2 \geq \alpha^2$. Hence,

$$
\mathbb{P}(||k^{-1}XR - k^{-1}X'R||_F^2 \geq \alpha^2) \leq \mathbb{P}\left(B^2 k^{-2} \sum_{j=1}^{k} \sum_{p=1}^{d} R_{p,j}^2 \geq \alpha^2\right) = \mathbb{P}(B^2 k^{-2} Y \geq \alpha^2)
$$

Therefore,

$$\begin{aligned}
\mathbb{P}(||k^{-1}XR - k^{-1}X'R||_F \geq \alpha) &\leq \mathbb{P}(B^2k^{-2}Y \geq \alpha^2) \\
&= \mathbb{P}\left(Y \geq \left(\frac{\alpha k}{B}\right)^2\right) \\
&\leq \frac{\mathbb{E}[\exp{(Y)}]}{\exp{\left(\left(\frac{\alpha k}{B}\right)^2\right)}} \\
&= \frac{((1-p_0)(e-1)+1)^{dk}}{\exp{\left(\left(\frac{\alpha k}{B}\right)^2\right)}}
\end{aligned}$$

where the second inequality follows by the Markov bound, and the last equality follows from the moment generating function of the binomial distribution. Plugging in the bound for α in the proposition yields the claim.

□

Now, set $p_0 = \frac{1}{3}$ and $\beta = \delta/2$ in Proposition 1. Then,

$$\alpha \geq \frac{B}{\sqrt{k}}\sqrt{d\ln((2/3)(e-1)+1) - k^{-1}\ln(\delta/2)}$$

With this probabilistic sensitivity bound, we use the following lemma from Kenthapadi et. al. [28] to show that Steps 1-3 of Algorithm 1 satisfy $(B, \epsilon_1, \delta_1)$-TDP.

Lemma 6. *[Lemma 1 from [28]; Lemma 3 from [30]] The mechanism given by $M(X) = f(X) + G$ satisfies (B, ϵ, δ)-TDP if $\delta < \frac{1}{2}$ and $G \in \mathbb{R}^{n\times k}$ with each $G_{p,j} \sim_{i.i.d} N(0, \sigma^2)$, where*

$$\sigma^2 = 2\Delta_2(f)^2\frac{\ln(1/(2\delta)) + \epsilon}{\epsilon^2}$$

and $\Delta_2(f)$ is the global sensitivity of f.

Proposition 2. *Set $p_0 = \frac{1}{3}$ and $\beta = \delta/2$ in the projection distribution of Proposition 1. Fix a matrix R generated according to this distribution. Then the mechanism given by $M(X) = XR + G$ satisfies (B, ϵ, δ)-TDP if $\delta < \frac{1}{2}$ and $G \in \mathbb{R}^{n\times k}$ with each $G_{p,j} \sim_{i.i.d} N(0, \sigma^2)$, where*

$$\sigma = \frac{B}{\sqrt{k}}\sqrt{d\ln((2/3)(e-1)+1) - k^{-1}\ln(\delta/2)}\frac{\sqrt{2(\ln(1/\delta) + \epsilon)}}{\epsilon}$$

Proof. By the Proposition 1, with probability $1-\beta$, $||k^{-1}XR - k^{-1}X'R||_F \leq \alpha$. Set $\beta = \frac{\delta}{2}$. Now, take $f(X) = k^{-1}XR$. Then $\Delta_2(f) = ||k^{-1}XR - k^{-1}X'R||_F \leq \alpha$ with probability $1 - \frac{\delta}{2}$. Plugging in the value of α in Lemma 6 with $\delta/2$ yields the claim. □

Next, we show that Step 4 of Algorithm 1 satisfies $(B, \epsilon_2, \delta_2)$-TDP.

Lemma 7. *Let C be the covariance matrix of $X \in \mathbb{L}_2^{n\times d}$. Then the mechanism $C + G$ satisfies (B, ϵ, δ)-TDP where $G_{i,j} \sim_{i.i.d} N(0, 2g_C^2\ln(1.25/\delta)/\epsilon^2)$ for all $i \geq j$, and $G_{i,j} = G_{j,i}$ for all $i < j$, where $g_C \leq 2B$ is the worst-case Frobenius norm for the difference of covariance matrices that are B-neighbors, provided $\epsilon \in (0,1)$.*

Proof. Suppose X and Z are B-neighbors. Then there exists a unique row i such that $||X_i - Z_i||_2 \leq B$, with all other rows being identical between X and Z. For notational simplicity, denote $X_i = x$ and $Z_i = z$. The sensitivity of the covariance computation is $||X^T X - Z^T Z||_F = ||x^T x - z^T z||_F$. Then

$$\begin{aligned}
||X^T X - Z^T Z||_F &= ||x^T x - z^T z||_F \\
&= ||(x^T x - x^T z) + (x^T z - z^T z)||_F \\
&= ||x^T (x - z) + (x - z)^T z||_F \\
&\leq ||x^T (x - z)||_F + ||(x - z)^T z||_F \\
&= ||x||_2 ||x - z||_2 + ||x - z||_2 ||z||_2 \\
&\leq 2B
\end{aligned}$$

Hence $g_C \leq 2B$. Since $\epsilon \in (0,1)$, we can use this value as the sensitivity in the Gaussian perturbation mechanism [21, 80, 25]. The claim now follows by the same line of reasoning as Theorem 2 of Dwork et al. [65]. □

Taken together, the results of this section all culminate to the following theorem.

Theorem 2. *Algorithm 1 satisfies $(B, \epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-TDP.*

Proof. By the Proposition 2, Steps 1-3 of Algorithm 1 satisfy $(B, \epsilon_1, \delta_1)$-TDP. By Lemma 7, Step 4 satisfies $(B, \epsilon_2, \delta_2)$-TDP. By the post-processing lemma (Lemma 1), Step 5 does not increase the privacy loss from Step 4. Step 6 combines the computed values, so by the sequential composition lemma (Lemma 2), our algorithm satisfies $(B, \epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-TDP. □

2 Supplementary Note: Additional Information on Privacy Measures

2.1 Relative protection score using the holdout-approach for attribute inference

In this section, we provide additional technical details on our quantification of attribute inference protection. We follow the approach presented in Giomi et al. [72]. We consider an adversary who has access to the privatized data X_{priv}, the privatization algorithm A, and an auxiliary information source: a column-wise submatrix S of $1 \leq h < d$ columns of X. That is, the adversary has access to h columns of the original data on all n individuals. Using these three objects, the adversary's goal is to infer the missing $d - h$ columns of X.

Our measurement of attribute inference protection is adopted from Giomi et al. [72], which uses a holdout-approach similar to machine learning applications. For completeness and self-contained exposition, we describe the measurement procedure (and provide additional details) below.

For a column j that an adversary seeks to infer in a database X, let $p(Z; S)_j$ represent Z's *protection score for column j* when S is known. As noted in *Methods* in the main text, this the proportion of individuals whose values the adversary correctly infers in column j (up to a tolerance of 5% error) using a privatized database Z and auxiliary info S.

Given the original data X, randomly select 500 individuals to remove from X. Denote X without these 500 individuals as W (working-set) and the dataset with these 500 individuals as H (holdout-set). Run the privatization algorithm A on W to construct W_{priv}. Run the attribute inference program procedure on both W_{priv} and H to generate $p(W_{priv}; S)_j$ and $p(H; S)_j$.

For a column j, the protection score $p(W_{priv}; S)_j$ in isolation is difficult to interpret. For instance, if $p(W_{priv}; S)_j$ is small, then this could be attributed to multiple culprits. In one case, A preserved population-level information in W well without retaining "too much information" about any single individual; since the population level information is preserved, individual values could nonetheless be inferred. On the other hand, A could have retained "too much information" about an individual, enabling the inference of their values. In the first example, the success rate misidentifies the "statistical utility" of our privatized dataset with inference risk, whereas in the second example the success rate correctly identifies inference risk.

In order to delineate between these two cases, we compare $p(W_{priv}; S)_j$ with $p(H; S)_j$. Since W_{priv} was created using W and not H, if $p(W_{priv}; S)_j$ and $p(H; S)_j$ are close, then this can be ascribed to the algorithm A capturing statistical information from the whole working population W and not from a particular individual. Alternatively, if $p(W_{priv}; S)_j << p(H; S)_j$, then it likely has learned "too much" about particular individuals. Thus, we define the *relative protection score of column j under S* as[3]

$$r_{j,S} = \frac{p(W_{priv}; S)_j}{p(H; S)_j}$$

2.2 Computing the distinguishing protection score for our algorithm

Our privacy measure of distinguishing protection is rooted in the *privacy loss random variable*, a unitless quantity that measures the differences in an algorithm's behavior between classic neighbors $X, X' \in \mathbb{L}_2^{n \times d}$ [80]. The privacy loss random variable quantifies the adversary's advantage during the distinguishing game when a randomized algorithm A outputs X_{priv} (i.e., the ability to discern whether X or X' was more likely to produce X_{priv} under A) [84].

For Gaussian perturbation mechanisms with variance σ^2, the privacy loss random variable follows a Gaussian distribution with mean $||f(X) - f(X')||_F^2/(2\sigma^2)$ [25]. Therefore, the worst-case expected privacy loss is given by $\sup||f(X) - f(X')||_F^2/(2\sigma^2)$ where the supremum is taken over X and X' classic neighbors.

Algorithm 1 is comprised of two Gaussian perturbation mechanisms in Steps 3 and 4. Let E_3 and E_4 represent the worst-case expected privacy loss induced by Steps 3 and 4. Computing these quantities from Proposition 2 and Lemma 7 yields

$$E_3 = \frac{\epsilon_1^2}{4(\ln(1/\delta_1) + \epsilon_1)} \text{ and } E_4 = \frac{\epsilon_2^2}{16\ln(1.25/\delta_2)}$$

Since E_3 and E_4 are defined for classic neighbors, they do not contain information about B. To enable comparison of our algorithm across all values of B, we use the switch lemma to translate all (B, ϵ, δ) to $(2, \hat{\epsilon}, \hat{\delta})$, where $s = \lceil 2B^{-1} \rceil$, $\hat{\epsilon} = s\epsilon$, and $\hat{\delta} = \min\{1, \frac{e^{s\epsilon}-1}{e^{\epsilon}-1}\delta\}$. Hence, an upper bound on the worst-case expected privacy loss of our algorithm with parameters $B, \epsilon_1, \delta_1, \epsilon_2, \delta_2$ is given by

$$\mathcal{U} = \frac{\hat{\epsilon}_1^2}{4(\ln(1/\hat{\delta}_1) + \epsilon_1)} + \frac{\hat{\epsilon}_2^2}{16\ln(1.25/\hat{\delta}_2)} \in [0, \infty) \cup \{\infty\}$$

To facilitate consistent interpretation with singling-out and attribute inference protection (each of which take values in $[0, 1]$), we define the distinguishing protection as $\mathcal{D} = (\mathcal{U} + 1)^{-1} \in [0, 1]$.

3 Supplementary Figures

[3] Giomi et al. define this an equivalent quantity in terms of a *risk score*, which is $1 - r_{j,S}$.

Table 1: Summary statistics of the datasets used in our empirical analyses.

	Togo	**Nigeria**
Number of Subscribers	4,201	20,788
Number of Features	10	15
Targeting Variable	Consumption	Loan Repayment
Targeting Criteria	$\leq 29^{th}$ percentile	Low-risk borrower (1)
Range of Targeting Variable	[0.005, 1]	{0,1}
Mean of Targeting Variable	0.073	0.633
SD of Targeting Variable	0.068	0.482
Skewness of Targeting Variable	4.207	-0.551